Kernel-Based Distribution Features for Statistical Tests and Bayesian Inference

Wittawat Jitkrittum

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy
of
University College London.

Gatsby Computational Neuroscience Unit
University College London

November 2017

Declaration

I, Wittawat Jitkrittum, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Wittawat Jitkrittum

Abstract

The kernel mean embedding is known to provide a data representation which preserves full information of the data distribution. While typically computationally costly, its nonparametric nature has an advantage of requiring no explicit model specification of the data. At the other extreme are approaches which summarize data distributions into a finite-dimensional vector of hand-picked summary statistics. This explicit finite-dimensional representation offers a computationally cheaper alternative. Clearly, there is a trade-off between cost and sufficiency of the representation, and it is of interest to have a computationally efficient technique which can produce a data-driven representation, thus combining the advantages from both extremes.

The main focus of this thesis is on the development of linear-time mean-embedding-based methods to automatically extract informative features of data distributions, for statistical tests and Bayesian inference. In the first part on statistical tests, several new linear-time techniques are developed. These include a new kernel-based distance measure for distributions, a new linear-time nonparametric dependence measure, and a linear-time discrepancy measure between a probabilistic model and a sample, based on a Stein operator. These new measures give rise to linear-time and consistent tests of homogeneity, independence, and goodness of fit, respectively. The key idea behind these new tests is to explicitly learn distribution-characterizing feature vectors, by maximizing a proxy for the probability of correctly rejecting the null hypothesis. We theoretically show that these new tests are consistent for any finite number of features.

In the second part, we explore the use of random Fourier features to construct approximate kernel mean embeddings, for representing messages in expectation propagation (EP) algorithm. The goal is to learn a message operator which predicts EP outgoing messages from incoming messages. We derive a novel two-layer random feature representation of the input messages, allowing online learning of the operator during EP inference.

For His Majesty King Bhumibol Adulyadej of Thailand
(1927-2016)

Acknowledgements

I would like to express my deep gratitude to my supervisor, Arthur Gretton, for his excellent supervision throughout my study. Arthur has always patiently listened to and helped solve problems I encountered. I thank the Gatsby Charitable Foundation for the financial support of my PhD study. I am grateful to Mijung Park and Zoltán Szabó for being a great collaborator over the years. I thank Zoltán Szabó's patience and kindness for mentoring me. I am grateful to Kenji Fukumizu for hosting me when I visited the Institute of Statistical Mathematics, and for fruitful research discussions. I thank all Gatsby Unit's members for intellectually stimulating discussions in the afternoon tea times. I thank my officemates: Carsen Stringer, Vincent Adam, Federico Mancinelli, Sanjeevan Ahilan, Kevin Wenliang Li, and Wenkai Xu, who have listened and responded to my complaints, and chatted about random scientific ideas with me. I thank Barry Fong for providing interesting stimulating questions from the perspective of a non-researcher. I thank my family for their love and support. A special thank goes to Jiamin Su who provided me emotional support in my final year.

Contents

Chapter 1

Introduction

We address the problem of finding computationally efficient, and informative *features* of data distributions, while making as few assumptions as possible on the underlying generating distributions. We consider two different contexts: 1) nonparametric statistical tests for comparing distributions, and 2) approximate Bayesian inference.

1.1 Informative Features for Statistical Tests

The task of nonparametric comparison of distributions is broad, and encompasses long-standing problems such as two-sample testing, independence testing, and goodness-of-fit testing. These will be the three problems that we consider in the first part. The three problems can be rephrased as finding the differences between two distributions P and Q by testing the null hypothesis $H_0\colon P = Q$ against the alternative $H_1\colon P \neq Q$ for some distributions P and Q. Two-sample testing aims to test $H_0\colon P = Q$ on the basis of two samples drawn from the two unknown distributions P and Q. Independence testing tests statistical dependence of two random vectors X and Y. Equivalently, this can be cast as testing $H_0\colon P_{xy} = P_x P_y$ using only the joint sample drawn from the unknown joint distribution P_{xy}, where P_x and P_y are the respective marginal distributions of X and Y. Goodness-of-fit testing examines whether a given sample follows a known probability distribution (model): given a sample from an unknown distribution Q, and a known model P, it tests $H_0\colon P = Q$. The knowledge of P is what distinguishes goodness-of-fit testing from the two-sample testing.

We consider only nonparametric testing, meaning that the assumptions made on the distributions P and Q are mild. Importantly, we do not assume any parametric family to which P and Q belong, in any of the three problems. By contrast, the well known t-test can be seen as a form of restrictive two-sample test, where the two samples are represented by their empirical means. In this case, the difference between P and Q can be detected only when there is a difference between the means, a strong implicit assumption which may not hold in practice. Many modern nonparametric tests are based on the use of positive definite kernels whose corresponding reproducing kernel Hilbert spaces (RKHSs) are of infinite dimensional [Gretton et al., 2005b,c, Moulines et al., 2008, Gretton et al., 2012a, Chwialkowski et al., 2014, Chwialkowski

and Gretton, 2014, Chwialkowski et al., 2015, 2016]. The central idea is based on the representation of empirical distributions with the so called *kernel mean embeddings* [Berlinet and Thomas-Agnan, 2004, Smola et al., 2007].

Given a distribution P, its mean embedding is defined as the expectation of the feature map implicitly defined by the kernel, resulting in a representation of P as a point in the RKHS. It can shown that if the kernel is *characteristic* [Sriperumbudur et al., 2011], then the mean map is injective, so that the distance between two distributions as measured in the embedded RKHS defines a distance in the original space of distributions. The use of such RKHS distance has led to the maximum mean discrepancy (MMD) test [Gretton et al., 2006, 2012a], a modern state-of-the-art nonparametric two-sample test. Measuring the distance between P_{xy} and $P_x P_y$ by the MMD leads to the so called Hilbert Schmidt Independence Criterion (HSIC) [Gretton et al., 2005a,b] which can be used to construct a consistent independence test. For goodness-of-fit testing, to the best of our knowledge, a multivariate, nonparametric (i.e., the model P is not restricted to a parametric family) test has not been studied in the literature until the recent works of Liu et al. [2016] and Chwialkowski et al. [2016]. The tests in these two works rely on a kernelized Stein operator to define the test statistic. Briefly, these tests rely on a test statistic given by the empirical expectation under Q of a P-dependent function (constructed by a Stein operator of P). It was shown that asymptotically the expectation of such a function is zero if and only if the sample follows P, allowing one to conduct a nonparametric goodness-of-fit test without the need of a sample from P. Further, the dependency on P in the constructed function is only through the gradient (with respect to the input variable) of the log density. This means that the normalizer of P does not need to be known, and the test can be applied to a complex model whose normalizer may be computationally intractable.

Motivations and Contributions Despite strong theoretical properties, a bottleneck common to all the kernel-based tests is their high runtime complexity, which is quadratic in the sample size. This high cost means that these tests can be applied to only problems of small size. In the followings, we briefly describe three commonly used techniques to reduce the computation.

1. **Random Fourier Features** [Rahimi and Recht, 2007, Zhao and Meng, 2014, Zhang et al., 2017]: this approach aims to approximate evaluation of the kernel with a dot product in a finite-dimensional space, constructed by randomly sampling from the spectral density of the kernel. The test statistic can then be rewritten in its "primal form" so that the dominant term in the computational complexity is the number of features, rather than the sample size. While this approach allows the test to be applied to larger problems, the use of a finite-dimensional kernel implies that the test is no longer consistent. That is, since only finitely many statistics can be captured, there exists a pair of P, Q that cannot be distinguished by the test. To ensure test consistency, as sample size increases, a growing number of features is needed. However, this defeats the

purpose of reducing the runtime in the first place.

2. **Incomplete Cholesky Factorization and Nyström Method** [Williams and Seeger, 2001, Zhang et al., 2017]: incomplete Cholesky factorization and the Nyström method approximate the kernel Gram matrix with a low-rank factorization by taking advantage of the fact that the kernel has a rapidly decaying spectrum (e.g., the Gaussian kernel). The test statistic can be rewritten in terms of the low-rank factors of the Gram matrix, thus reducing the computation. The asymptotics and consistency of a test based on the Nyström method or incomplete Cholesky factorization remain a challenging open question. Further, when the kernel spectrum does not decay sufficiently rapidly, the reduced rank may still need to be large to accurately approximate the Gram matrix.

3. **Incomplete U-Statistic** [Gretton et al., 2012a, Zaremba et al., 2013, Zhang et al., 2017]: many kernel-based test statistics can be written as a second-order U-statistic (Section A.1: U-Statistics), taking the form $T_n = \frac{2}{n(n-1)} \sum_{i<j} h(\mathbf{z}_i, \mathbf{z}_j)$ for some function h, where n is the sample size, and $\{\mathbf{z}_i\}_{i=1}^n$ is the observed sample. Computational complexity of T_n is $\mathcal{O}(n^2)$. The idea of an incomplete U-statistic is to subsample summands in T_n so that the number of terms left is of order $\mathcal{O}(n)$. The result is a test statistic which can be computed in linear-time, and is still unbiased. An advantage over the previous two approaches is that test consistency still remains (if the original T_n yields a consistent test). A disadvantage is that it tends to give a test with low test power (i.e., the probability of rejecting H_0 when it is false) for finite n, due to the increase in the variance of the statistic.

An equally pressing issue is the choice of the kernel itself. Of all the aforementioned kernel tests, apart from the MMD two-sample test for which kernel optimization has been investigated [Gretton et al., 2012b, Sutherland et al., 2016], there is no principled way of optimizing kernels in the HSIC test of independence, and the kernelized Stein test of goodness of fit. These are the motivations for our proposals. Our goals are to develop new kernel tests for the three types of testing which address the drawbacks of existing kernel tests, while maintaining their advantages:

1. The new tests have a principled way of choosing kernels and all other hyperparameters.
2. The new tests, including the parameter tuning procedure, run in linear-time (with respect to the sample size).
3. The new tests can be used with multivariate random variables, are nonparametric and consistent.

The key idea behind these new tests is to learn explicit features (points in the same domain as the input data) so as to maximize the rate of detecting the differences between the two distributions i.e., the test power. The features not only allow one to

avoid the expensive (quadratic in the sample size) computation of the distance in the RKHS, but turn out to also pinpoint and indicate where the two distributions differ. For instance, in the goodness-of-fit test, the latter property makes the test interpretable: it uniquely gives evidence indicating the region in which the model P fails to fit the data. All hyperparameters can be automatically tuned so as to maximize the (lower bound on the) test power. This series of works thus simultaneously addresses a number of long-standing issues in kernel-based hypothesis testing, namely, unavailability of a parameter tuning procedure, and high runtime complexity. Importantly, our linear-time tests are consistent for *any* finite number of features. To reiterate, the commonly used random Fourier features [Rahimi and Recht, 2007] to speed up kernel-based tests requires a growing number of features to guarantee test consistency.

Our new two-sample test, independence test, and goodness-of-fit test are described in Chapter 3, Chapter 4, and Chapter 5, respectively.

1.2 Informative Features for Learning to Infer

Given a graphical model consisting of *factors* (non-negative functions e.g., conditional probability densities) capturing how neighboring variables interact, the goal of Bayesian inference is to infer the posterior distribution of some variables of interest, conditioning on observed realizations of others. A commonly used approximate inference scheme is expectation propagation (EP), which recursively passes evidence from the observed variables in the form of outgoing messages (i.e., functions or distributions), to the variables to be inferred. Each message can be computed locally at a factor as a function of incoming messages from the neighbors.

Goals A major challenge is that computing a message typically involves an intractable integral over a complicated factor, and may require an expensive numerical integration. A typical approach is to manually compute (or approximate) the integral for each considered factor, and implement accordingly in the inference engine. This is the approach taken by Infer.NET [Minka et al., 2014], a probabilistic programming framework which supports EP. While the modeler can freely compose their model from supported factors known to the inference engine, using a customized factor still requires manual implementation of the outgoing messages. Our goal is to automate the computation of outgoing messages in EP for arbitrary factors, so that no manual derivation is needed. Further, we also require that the overhead resulting from such automation be kept minimal.

An approach due to Barthelmé and Chopin [2011] is to compute the messages via importance sampling. Although this approach eliminates the need of manual derivation of message computation, it suffers from high computational complexity. Heess et al. [2013] use neural networks to predict outgoing messages from incoming messages, replacing the expensive numerical integration. The neural networks are trained offline on importance sampled instances of incoming/outgoing message pairs. While there is a large gain in runtime, this approach requires training data that cover

the relevant regions of input messages to be encountered during the inference. In practice, such relevant regions are unknown during the training phase. Eslami et al. [2014] address this problem by considering online learning (during the EP inference) of a random forest based message predictor. Whenever the random forest decides it is uncertain of the input messages, it queries the importance sampling oracle for the outgoing message, and updates itself. Otherwise, the outgoing message is efficiently predicted by the random forest. A disadvantage of random forests is that uncertainty estimation relies on unproven heuristics, and is highly non-smooth. This means that the random forest can be uncertain of even a tuple of input messages which is in the region populated by the training messages. As a result, the importance sampling oracle is queried more frequently than necessary.

In our collaborative work [Jitkrittum et al., 2015] with Google DeepMind, we use the kernel mean embedding to represent incoming messages, and construct a Gaussian process regression function that learns to send EP messages. Two-layer random Fourier features for distributional input are developed and used to further improve the speed, making the overall complexity linear in the sample size. The result is an automatic inference engine (implemented in Infer.NET [Minka et al., 2014]) which quickly learns to send messages for any factor that can be sampled, eliminating the need of manually deriving message computation. We show empirically that the inference quality matches that of Infer.NET which relies on handcrafted factors. This work addresses simultaneously both the model expressiveness (i.e., automatically handling complicated factors) and computational tractability. The developed tool can be useful in probabilistic programming, since it makes inference fast and practical for complex models. This study is described in Chapter 6.

1.3 Structure of the Thesis

We start in Chapter 2 with some brief background on kernel methods for learning on distributions. Chapter 3 describes the new linear-time two-sample test, which relies on explicit difference-characterizing features. The technique developed in this chapter is further extended in Chapter 4 to construct a new linear-time independence test, where the learned features indicate the regions in which the joint distribution and the product of the marginal distributions differ most. In Chapter 5, the test based on the kernelized Stein operator is discussed and extended, leading to a new linear-time goodness-of-fit test that also gives features indicating where the model fails to fit the data. A brief background on Bayesian inference is given in the beginning of Chapter 6, followed by our contribution to automate the expectation propagation algorithm. This is possible due to the developed two-staged Fourier feature representation of the incoming messages. We end the thesis with conclusions and remarks on a number of potential future studies in Chapter 7.

The four main thesis chapters are based on the following publications which were published over the course of this thesis. Source code for all proposed methods in this

thesis is publicly available.

1. Chapter 3: Informative Features for Distinguishing Distributions

   W. Jitkrittum, Z. Szabó, K. Chwialkowski, A. Gretton. Interpretable Distribution Features with Maximum Testing Power. In NIPS, 2016. (Oral presentation)
   Code: https://github.com/wittawatj/interpretable-test

2. Chapter 4: Informative Features for Dependence Detection

   W. Jitkrittum, Z. Szabó, A. Gretton. An Adaptive Test of Independence with Analytic Kernel Embeddings. In ICML, 2017.
   Code: https://github.com/wittawatj/fsic-test

3. Chapter 5: Informative Features for Model Criticism

   W. Jitkrittum, W. Xu, Z. Szabó, K. Fukumizu, A. Gretton. A Linear-Time Kernel Goodness-of-Fit Test. In NIPS, 2017. (Oral presentation)
   Code: https://github.com/wittawatj/kernel-gof

4. Chapter 6: Informative Features for Automated Expectation Propagation

   W. Jitkrittum, A. Gretton, N. Heess, S. M. A. Eslami, B. Lakshminarayanan, D. Sejdinovic, Z. Szabó. Kernel-Based Just-In-Time Learning for Passing Expectation Propagation Messages. In UAI, 2015.
   Code: https://github.com/wittawatj/kernel-ep

Other Contributions Works published over the course of this thesis that are not included are

- M. Park,* **W. Jitkrittum**,* D. Sejdinovic. K2-ABC: Approximate Bayesian Computation with Infinite Dimensional Summary Statistics via Kernel Embeddings. In AISTATS, 2016. (*Equal contribution. Oral presentation.)
  Code: https://github.com/wittawatj/k2abc
- K. Iigaya, A. Jolivald, **W. Jitkrittum**, I. Gilchrist, P. Dayan, E. Paul, M. Mendl. Cognitive Bias in Ambiguity Judgements: Using Computational Models to Dissect the Effects of Mild Mood Manipulation in Humans. PLOS ONE, 2016.
- M. Park, **W. Jitkrittum**, A. Qamar, Z. Szabó, L. Buesing, M. Sahani. Bayesian Manifold Learning: The Locally Linear Latent Variable Model. In NIPS, 2015.
  Code: https://github.com/mijungi/lllvm

Chapter 2

Kernel Methods for Learning on Distributions

The use of kernel mean embeddings to measure the distance between two distributions is at the core of all our contributions. In this chapter, we provide a brief review of the theory of reproducing kernel Hilbert spaces (RKHSs), and kernel mean embeddings. In the following chapters, we will assume the background knowledge described in this chapter. For rigorous treatment of the theory of RKHSs, see Berlinet and Thomas-Agnan [2004], Steinwart and Christmann [2008]. A broad overview of kernel methods can be found in Muandet et al. [2017].

2.1 Reproducing Kernel Hilbert Space

There are a few equivalent definitions of a reproducing kernel Hilbert space (RKHS). We will present the simplest and the most suitable for our purpose. We first give the definition of a reproducing kernel.

Definition 2.1 (Reproducing kernel [Berlinet and Thomas-Agnan, 2004, Definition 1])**.** Let $\mathcal{F}$ be a Hilbert space of real-valued functions defined on a non-empty set $\mathcal{X}$. Write $\langle \cdot, \cdot \rangle$ to denote the inner product associated with $\mathcal{F}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be a *reproducing kernel* of $\mathcal{F}$ if

1. for all $\mathbf{x} \in \mathcal{X}$, $k(\mathbf{x}, \cdot) \in \mathcal{F}$,
2. for all $\mathbf{x} \in \mathcal{X}$ and for all $f \in \mathcal{F}$, $\langle f, k(\mathbf{x}, \cdot) \rangle = f(\mathbf{x})$.

The second property is known as the *reproducing property*. We will write $\mathcal{F}(k)$ for $\mathcal{F}$ when the dependency on k needs to be emphasized.

When the space $\mathcal{F}$ on which the inner product $\langle \cdot, \cdot \rangle$ is defined needs to be emphasized, we will write $\langle \cdot, \cdot \rangle_{\mathcal{F}}$.

Definition 2.2 (RKHS [Steinwart and Christmann, 2008, Section 4.2])**.** A reproducing kernel Hilbert space (RKHS) is a Hilbert function space $\mathcal{F}$ with a reproducing kernel k (as defined in Definition 2.1).

Existence of an RKHS does not require other conditions on the domain $\mathcal{X}$ besides that it be non-empty. In an RKHS, convergence in norm implies pointwise convergence. That is, given f and g in an RKHS $\mathcal{F}(k)$,

$$|f(\mathbf{x}) - g(\mathbf{x})| = |\langle f - g, k(\mathbf{x}, \cdot)\rangle_{\mathcal{F}}| \overset{(a)}{\leq} \|f - g\|_{\mathcal{F}} \|k(\mathbf{x}, \cdot)\|_{\mathcal{F}} = \|f - g\|_{\mathcal{F}} \sqrt{k(\mathbf{x}, \mathbf{x})}, \tag{2.1}$$

where (a) follows from the Cauchy-Schwarz inequality. The inequality in (2.1) implies that if f converges to g in the RKHS norm, then $f(\mathbf{x}) = g(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. The expression $k(\mathbf{x}, \cdot)$ can be interpreted in two different ways. Firstly, it can be seen as a function $\mathbf{v} \mapsto k(\mathbf{x}, \mathbf{v})$ constructed by fixing one argument of k to $\mathbf{x}$. Secondly, $k(\mathbf{x}, \cdot)$ can be seen as a vector in $\mathcal{F}$ (recall that $\mathcal{F}$ is a vector space). The reproducing property means that the evaluation of $f(\cdot)$ at $\mathbf{x}$ (i.e., $f(\mathbf{x})$) is given by the inner product between a feature vector $k(\mathbf{x}, \cdot)$ of $\mathbf{x}$, and a feature representation of the function $f(\cdot)$, which is denoted by f. This second interpretation means that $f \in \mathcal{F}$ can be seen as a parameter vector of the function $\mathbf{x} \mapsto \langle f, k(\mathbf{x}, \cdot)\rangle$, and consequently $\mathcal{F}$ is a space of parameter vectors which can be used to define real-valued functions. We interchangeably write $f(\cdot)$ (the function itself) and f (the feature representation of the function) when the distinction is not important.

To illustrate the reproducing kernel, let us consider a simple concrete example. Let $\mathcal{X} = \mathbb{R}$, $\phi(x) := (x, x^2)^\top$, and define

$$k(x, y) := \langle \phi(x), \phi(y)\rangle_{\mathbb{R}^2} = xy + x^2 y^2, \tag{2.2}$$

where $\langle \cdot, \cdot \rangle_{\mathbb{R}^2}$ is the standard dot product in $\mathbb{R}^2$. In this case, the space $\mathcal{F}$ is the set of functions $\left\{ x \mapsto \sum_{i=1}^2 \alpha_i \phi_i(x) \mid \alpha_1, \alpha_2 \in \mathbb{R} \right\}$. Alternatively, since (α_1, α_2) fully specifies a function (i.e., a function's parameters), one can see $\mathcal{F} = \mathbb{R}^2$. It follows that the reproducing kernel $k(x, \cdot) = \phi(x)$. The function $x \mapsto \phi(x)$ is called the *canonical feature map* [Steinwart and Christmann, 2008, Lemma 4.19], or simply feature map. It can be seen that the reproducing property as described in Definition 2.1 holds.

Positive Definite Kernel In (2.2), we start with a feature map and define the kernel. In general, one can define k, a real-valued function of two arguments, directly without explicitly specifying the underlying canonical feature map. The key question is: what conditions are required for the function to be a reproducing kernel of some Hilbert space? The answer and related important properties are summarized in Theorem 2.4.

Definition 2.3 (Positive definite function [Steinwart and Christmann, 2008, Definition 4.15])**.** A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *positive definite* if for all $n \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and all $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$, we have $\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \geq 0$.

Theorem 2.4 (Positive definite kernel and RKHSs)**.** *Assume that k: $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite (see Definition 2.3). The following statements hold.*

1. *There exists a map* $\phi\colon \mathcal{X} \to \mathcal{F}$ *(not unique) such that, for all* $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$*, we have* $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{F}}$.

2. *(Moore-Aronszajn theorem) There is a unique Hilbert space* $\mathcal{F}$ *of functions on* $\mathcal{X}$ *for which* k *is a reproducing kernel.*

3. *If a Hilbert space of functions on a non-empty set* $\mathcal{X}$ *has a reproducing kernel, then it is unique [Steinwart and Christmann, 2008, Theorem 4.20].*

Henceforth, we will refer to a function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which is positive definite simply as a *kernel*. Theorem 2.4 implies that the key to construct an RKHS is to have a positive definite kernel. In general, the domain $\mathcal{X}$ can be a subset of a non-Euclidean space. There are kernels for graphs, text, strings, and even probability distributions [Shawe-Taylor and Cristianini, 2004]. Note that although there is only one unique reproducing kernel associated with $\mathcal{F}$, the underlying feature map ϕ may not be unique. For example, the feature maps $\phi(x) := \left(\frac{x}{\sqrt{2}}, \frac{x}{\sqrt{2}}, x^2\right)^\top$ or $\phi(x) := (x, x^2, 0)^\top$ define the same kernel in (2.2). In general, a kernel may even be associated with an infinite-dimensional feature map.

A kernel whose underlying feature map is infinite-dimensional can be a powerful tool for learning. Many learning algorithms relying on a prediction function $f_{\boldsymbol{\theta}}(\mathbf{x})$ which is linear in the parameter vector $\boldsymbol{\theta}$ can be *kernelized*. This means that the same learning algorithm is applied in the feature space specified by a feature map ϕ, as implicitly induced by a kernel. If the underlying Hilbert space $\mathcal{F}$ is infinite-dimensional, the result is a powerful learning algorithm which uses an infinite basis expansion given by the infinite-dimensional map ϕ. Although $\mathbf{x} \mapsto \phi(\mathbf{x})$ cannot be directly computed and stored, it is typically the case that linear learning algorithms can be reformulated in such a way that the solution requires only evaluations of the inner product in the feature space $\mathcal{F}$. Since the inner product is given by the kernel k, the solution can be easily computed. Reformulating the problem so that the dependency on the infinite-dimensional feature map is only through its inner product is known as the *kernel trick*. A commonly used kernel corresponding to an infinite-dimensional Hilbert space is the Gaussian kernel, also known as the radial basis function (RBF) kernel

$$k(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|_2^2}{2\sigma^2}\right), \tag{2.3}$$

where $\mathcal{X} \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$, and $\sigma > 0$ is the kernel bandwidth. We note that different choices of σ^2 define different kernels, and hence different RKHSs. Successful learning algorithms which are based on the kernel trick include support vector machine [Cortes and Vapnik, 1995], kernel principal component analysis [Schölkopf et al., 1997], as well as the maximum mean discrepancy [Gretton et al., 2012a], a distance measure between two distributions (described in Section 2.2).

2.2 Kernel Mean Embedding

Kernel mean embedding is a technique to represent distributions as points in an RKHS. According to Berlinet and Thomas-Agnan [2004, p. 189], the idea was first studied in the years 1975-1980 by Denis Bosq and C. Guilbart. Let P be a probability measure on $\mathcal{X}$, and k be a kernel associated with RKHS $\mathcal{F}$. The *mean embedding* of P as induced by k is defined as

$$\mu_P := \mathbb{E}_{\mathbf{x}\sim P}[k(\mathbf{x}, \cdot)], \tag{2.4}$$

which is an element of $\mathcal{F}$ (if μ_P exists). In words, the mean embedding of the distribution P is the expectation under P of the canonical feature map. Conditions under which μ_P exists and is in $\mathcal{F}$ are summarized in Lemma 2.5.

Lemma 2.5 (Gretton et al. [2012a, Lemma 3], Sriperumbudur et al. [2010, Theorem 1], Smola et al. [2007]). *If* $\mathbb{E}_{\mathbf{x}\sim P}\sqrt{k(\mathbf{x},\mathbf{x})} < \infty$, *then* $\mu_P \in \mathcal{F}$ *and* $\mathbb{E}_{\mathbf{x}\sim P}[f(\mathbf{x})] = \langle f, \mu_P\rangle_{\mathcal{F}}$ *for all* $f \in \mathcal{F}$.

The condition $\mathbb{E}_{\mathbf{x}\sim P}\sqrt{k(\mathbf{x},\mathbf{x})} < \infty$ implies that the mean embedding exists only if the distribution is a member of $\mathcal{P}_k := \{P \in \mathcal{P} \mid \int_{\mathcal{X}} \sqrt{k(\mathbf{x},\mathbf{x})}\, \mathrm{d}P(\mathbf{x}) < \infty\}$ where $\mathcal{P}$ is the set of all Borel probability measures. If, however, k is bounded i.e., $\sup_{\mathbf{x}\in\mathcal{X}} k(\mathbf{x},\mathbf{x}) < \infty$, then mean embeddings are well defined for any $P \in \mathcal{P}$ [Sriperumbudur et al., 2010, Proposition 2]. For instance, this is the case for the Gaussian kernel in (2.3), which is bounded by 1. Embedding distributions to points in a Hilbert space allows one to use standard operations in the Hilbert space to manipulate them. One such operation is that, as seen in Lemma 2.5, the expectation under P of any function f in the RKHS defined by k can be computed by taking the inner product of μ_P and f. This property is convenient for many tasks.

Empirical Estimation Given an independent and identically distributed (i.i.d.) sample $\{\mathbf{x}_i\}_{i=1}^n \sim P$, and a kernel k, the mean embedding can be estimated straightforwardly with its plug-in estimator $\hat{\mu}_P := \frac{1}{n}\sum_{i=1}^n k(\mathbf{x}_i, \cdot)$ i.e., replace P in (2.4) by its empirical counterpart $\hat{P} := \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}$, where $\delta_{\mathbf{x}}$ denotes a Dirac measure at $\mathbf{x} \in \mathcal{X}$. By Bernstein's inequality in separable Hilbert spaces, $\|\hat{\mu}_P - \mu_P\|_{\mathcal{F}} = \mathcal{O}_P(n^{-1/2})$, where $\mathcal{O}_P$ denotes the stochastic big-oh. In other words, the empirical mean embedding is a $\sqrt{n}$-consistent estimator of μ_P in $\mathcal{F}$-norm [Tolstikhin et al., 2016]. In machine learning applications, P is often unknown, and only its sample is observed. The empirical mean embedding $\hat{\mu}_P$ is thus the quantity of interest in practice instead of μ_P. We next describe applications of mean embeddings.

2.3 Maximum Mean Discrepancy

A useful operation on embedded distributions which has applications in hypothesis testing is measuring their distance.[1] As before, let $k\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ be the kernel

[1] We use "distance" here as a generic, non-technical term. The formal concept of a mathematical distance will be referred to as a metric.

associated with the RKHS $\mathcal{F}$. The distance of two distributions P and Q defined on $\mathcal{X}$, as measured by their mean embeddings is known as the *maximum mean discrepancy* (MMD) [Gretton et al., 2006, 2012a]:

$$\begin{aligned} \mathrm{MMD}(P, Q) &:= \sup_{\|f\|_{\mathcal{F}} \leq 1} \mathbb{E}_{\mathbf{x} \sim P}[f(\mathbf{x})] - \mathbb{E}_{\mathbf{y} \sim Q}[f(\mathbf{y})] && (2.5) \\ &= \sup_{\|f\|_{\mathcal{F}} \leq 1} \langle \mu_P - \mu_Q, f \rangle_{\mathcal{F}} && (2.6) \\ &\overset{(a)}{=} \|\mu_P - \mu_Q\|_{\mathcal{F}}, && (2.7) \end{aligned}$$

where we use the reproducing property, and at (a) we use the fact that an inner product achieves its supremum when the two vectors are parallel. MMD as defined in (2.5) is an instance of an *integral probability metric* (IPM) [Müller, 1997], a pseudometric on the space of probability measures. A general IPM takes the form

$$\mathrm{IPM}(\mathcal{H}, P, Q) = \sup_{h \in \mathcal{H}} \mathbb{E}_{\mathbf{x} \sim P}[f(\mathbf{x})] - \mathbb{E}_{\mathbf{y} \sim Q}[f(\mathbf{y})], \tag{2.8}$$

and is parametrized by a class $\mathcal{H}$ of real-valued bounded measurable functions on $\mathcal{X}$. The choice $\mathcal{H}$ is crucial in making the IPM a metric (it is always a pseudometric regardless of $\mathcal{H}$). Choices of $\mathcal{H}$ which define a metric include $C_b(\mathcal{X})$, the space of bounded continuous functions on $\mathcal{X}$ [Dudley, 2002, Lemma 9.3.2], and the space of all functions on $\mathcal{X}$ that are bounded, Lipschitz [Shorack, 2000, p. 540, Definition 2.2]. The latter is known as the Dudley metric. More examples of IPMs can be found in Sriperumbudur et al. [2010, p. 1519]. The MMD considers $\mathcal{H} = \{f \mid \|f\|_{\mathcal{F}} \leq 1\}$, a unit ball in the RKHS $\mathcal{F}$.

Witness Function The RKHS function that attains the supremum in (2.6) is known as the *witness function* [Gretton et al., 2012a, Section 2.3]:

$$f^*(\mathbf{v}) \propto \mathbb{E}_{\mathbf{x} \sim P}[k(\mathbf{x}, \mathbf{v})] - \mathbb{E}_{\mathbf{y} \sim Q}[k(\mathbf{y}, \mathbf{v})], \tag{2.9}$$

which is proportional to the difference of the mean embeddings of P and Q. If we interpret the kernel k in (2.9) as a smoothing kernel for density estimation,[2] we see that the witness function is positive when the density of P exceeds the density of Q, and negative otherwise. The witness function can be used to visualize regions (in the domain $\mathcal{X}$) in which P and Q differ [Lloyd and Ghahramani, 2015] when the input dimension d is low. The RKHS norm of the witness function is the MMD.

2.3.1 Characteristic Kernels

As in IPMs, MMD is in general a pseudometric, unless k (and hence $\mathcal{F}$) is *characteristic* (Definition 2.6).

[2] In general a smoothing kernel and a positive definite kernel are two different objects. A smoothing kernel is a non-negative real-valued integrable function $K\colon \mathcal{X} \to \mathbb{R}$ such that $\int_{\mathcal{X}} K(u)\,\mathrm{d}u = 1$ (normalized), and $K(-u) = K(u)$.

Definition 2.6 (Characteristic kernels [Fukumizu et al., 2008, Sriperumbudur et al., 2011]). A kernel k is said to be characteristic if the mean map $P \mapsto \mathbb{E}_{\mathbf{x}\sim P}[k(\mathbf{x}, \cdot)]$ is injective on $\mathcal{P}$ (the set of all Borel probability measures). Equivalently, k is characteristic if $\mathrm{MMD}(P, Q) = \|\mu_P - \mu_Q\| = 0 \iff P = Q$ for any $P, Q \in \mathcal{P}$.

The injectivity of the mean map $P \mapsto \mathbb{E}_{\mathbf{x}\sim P}[k(\mathbf{x}, \cdot)]$ by the use of a characteristic kernel implies that distinct distributions are mapped to distinct points in the RKHS $\mathcal{F}$, allowing the distance in $\mathcal{F}$ (i.e., the MMD) to separate different distributions. Examples of characteristic kernels include the Gaussian kernel in (2.3), the Laplace kernel $k(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|_2}{\sigma}\right)$, the B-spline kernel, and the Matérn class of kernels [Sriperumbudur et al., 2010, Section 3.2].

An example of a non-characteristic kernel is the one given in (2.2) on $\mathbb{R} \times \mathbb{R}$. Since the feature map is $k(x, \cdot) = (x, x^2)^\top$, the mean map of a distribution P is $\mathbb{E}_{x\sim P}(x, x^2)^\top$, which captures only the first two moments of P (if exist). If P and Q share the same first two moments, and differ in higher-order moments, then $\mathrm{MMD}(P, Q) = 0$ even though $P \neq Q$. This kernel can be seen as a special case of the degree-D *polynomial kernel* $k_{c,D} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with $\mathcal{X} \subseteq \mathbb{R}^d$, given as

$$k_{c,D}(\mathbf{x}, \mathbf{y}) := (\mathbf{x}^\top \mathbf{y} + c)^D,$$

where $c \geq 0$ and $D \in \mathbb{N}$ are the two parameters of the kernel. It can be shown that the feature map $k_{c,D}(\mathbf{x}, \cdot)$ has dimensions indexed by monomials of $\mathbf{x}$. Specifically, $k_{c,D}(\mathbf{x}, \cdot) \in \mathbb{R}^{\binom{d+D}{D}}$ such that $k_{c,D}(\mathbf{x}, \cdot)_{\mathbf{a}} = \prod_{j=1}^d x_j^{a_j}$, $\mathbf{a} := (a_1, \ldots, a_d)$, and $\sum_{j=1}^d a_j \leq D$ [Shawe-Taylor and Cristianini, 2004, Section 9.1]. The polynomial kernel $k_{c,D}$ for any $c > 0$ and $D \in \mathbb{N}$ is not characteristic. Being characteristic is only one of many properties a kernel can have. Other useful properties will be discussed in Section 2.5.

2.3.2 Estimation, Convergence, and Asymptotic Distributions

By expanding the square of (2.7), we obtain

$$\begin{aligned}
\mathrm{MMD}^2(P, Q) &= \|\mu_P - \mu_Q\|_{\mathcal{F}}^2 \\
&= \langle \mu_P, \mu_P \rangle_{\mathcal{F}} + \langle \mu_Q, \mu_Q \rangle_{\mathcal{F}} - 2 \langle \mu_P, \mu_Q \rangle_{\mathcal{F}} \\
&= \mathbb{E}_{\mathbf{x}\sim P}\mathbb{E}_{\mathbf{x}'\sim P} k(\mathbf{x}, \mathbf{x}') + \mathbb{E}_{\mathbf{y}\sim Q}\mathbb{E}_{\mathbf{y}'\sim Q} k(\mathbf{y}, \mathbf{y}') - 2\mathbb{E}_{\mathbf{x}\sim P}\mathbb{E}_{\mathbf{y}\sim Q} k(\mathbf{x}, \mathbf{y}). \quad (2.10)
\end{aligned}$$

Given samples $\{\mathbf{x}_i\}_{i=1}^m \overset{i.i.d.}{\sim} P$ and $\{\mathbf{y}_i\}_{i=1}^n \overset{i.i.d.}{\sim} Q$, an unbiased estimator of (2.10) is given by

$$\widehat{\mathrm{MMD}^2}(P, Q) = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j\neq i} k(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j\neq i} k(\mathbf{y}_i, \mathbf{y}_j) - \frac{2}{mn} \sum_{i=1}^m \sum_{j=1}^n k(\mathbf{x}_i, \mathbf{y}_j), \quad (2.11)$$

which can be computed straightforwardly. Unlike other divergences or distances for distributions which require density estimation (e.g., an L_2 distance between Parzen window estimates), the MMD estimator in (2.11) can be directly computed given only

a kernel k and samples. Assume that $m = n$. Then, this unbiased MMD estimator is a one-sample second-order U-statistic (see Section A.1) where its U-statistic core is

$$h(\mathbf{z}_i, \mathbf{z}_j) := k(\mathbf{x}_i, \mathbf{x}_j) + k(\mathbf{y}_i, \mathbf{y}_j) - k(\mathbf{x}_i, \mathbf{y}_j) - k(\mathbf{x}_j, \mathbf{y}_i), \tag{2.12}$$

and $\mathbf{z}_i := (\mathbf{x}_i, \mathbf{y}_i) \overset{i.i.d.}{\sim} P \times Q$ [Gretton et al., 2012a, Lemma 6]. The unbiased estimator can then be written as

$$\widehat{\mathrm{MMD}^2}(P, Q) = \binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{j>i} h(\mathbf{z}_i, \mathbf{z}_j). \tag{2.13}$$

Thus, convergence results and asymptotic distributions can be obtained by appealing to the theory of U-statistics. Directly based on Lemma A.3 and Lemma A.4, the asymptotic distributions of the unbiased MMD estimator can be derived under two cases: when $P = Q$ and when $P \neq Q$. That is, when $P = Q$, $\widehat{\mathrm{MMD}^2}(P, Q)$ follows an infinite weighted sum of chi-squares as $n \to \infty$ [Gretton et al., 2012a, Theorem 12]. When $P \neq Q$, $\widehat{\mathrm{MMD}^2}(P, Q)$ is asymptotically normally distributed with the mean given by $\mathrm{MMD}^2(P, Q) > 0$ [Gretton et al., 2012a, Corollary 16].

A Linear-Time Estimator Assume that $m = n$. The cost for computing the unbiased MMD estimator in (2.11) is $\mathcal{O}(n^2)$ which is expensive for large n. Let $n_2 := \lfloor n/2 \rfloor$ where $\lfloor \cdot \rfloor$ is the floor function. A linear-time estimator was proposed in Gretton et al. [2012a, Lemma 14], and is given by

$$\widehat{\mathrm{MMD}}_1^2(P, Q) = \frac{1}{n_2} \sum_{i=1}^{n_2} h((\mathbf{x}_{2i-1}, \mathbf{y}_{2i-1}), (\mathbf{x}_{2i}, \mathbf{y}_{2i})), \tag{2.14}$$

where h is defined in (2.12). This estimator is an incomplete U-statistic [Janson, 1984], which considers a subset of the summands of (2.13). It can be seen that $\widehat{\mathrm{MMD}}_1^2(P, Q)$ is unbiased and can be computed in $\mathcal{O}(n)$ time. The central limit theorem implies that $\widehat{\mathrm{MMD}}_1^2(P, Q)$ is asymptotically normally distributed with the mean given by $\mathrm{MMD}^2(P, Q)$. Compared to the quadratic-time estimator, the linear-time estimator has higher variance.

2.4 Applications of Mean Embedding

Two-Sample Testing A natural application of the MMD is to use it as a test statistic for two-sample testing [Gretton et al., 2012a]. In two-sample-testing or test of homogeneity, given samples $\{\mathbf{x}_i\}_{i=1}^m \overset{i.i.d.}{\sim} P$ and $\{\mathbf{y}_i\}_{i=1}^n \overset{i.i.d.}{\sim} Q$, the goal is to test the null hypothesis $H_0 : P = Q$ against the alternative $H_1 : P \neq Q$ based on only the samples. The test is achieved by comparing the test statistic to a *test threshold*. If the statistic exceeds the threshold, the null hypothesis H_0 is rejected. A common choice of the test threshold is the $(1 - \alpha)$-quantile of the null distribution of the statistic i.e.., the distribution of the test statistic assuming that H_0 is true. The quantity α is known as the significance level of the test, and is predetermined in advance before observing the samples. This

choice for the test threshold means that the type-I error (false rejection of H_0 when it is true) will not exceed α.

When the quadratic-time MMD estimator in (2.11) is used, the asymptotic null distribution is given by an infinite weighted sum of chi-squared random variables, where the weights are the eigenvalues of an operator defined based on the kernel [Gretton et al., 2012a, Theorem 12]. The distribution does not have a closed-form expression. One way to estimate the $(1-\alpha)$-quantile of the null distribution is by the permutation testing, using the bootstrap on the aggregated samples [Arcones and Gine, 1992]. This procedure involves repeatedly computing the quadratic-time estimator on shuffled samples, and can be computationally expensive. An approximation to the intractable null distribution can be obtained by fitting Pearson curves to the first four moments of the MMD [Gretton et al., 2012a, Section 5]. Further alternatives include a consistent estimation of the eigenvalues from the spectrum of the Gram matrix, and fitting a Gamma distribution to the null distribution [Gretton et al., 2009]. The latter provides a fast procedure for determining the quantile, but is less accurate compared to other approaches. Two-sample testing will be discussed in detail in Chapter 3.

Independence Testing Let $X \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $Y \in \mathcal{Y} \subseteq \mathbb{R}^{d_y}$ be two multivariate random variables. Let P_{xy} be the joint distribution of (X, Y), and P_x, P_y be the respective marginal distributions of X and Y. Given a joint sample $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \overset{i.i.d.}{\sim} P_{xy}$, an independence test proposes the null hypothesis $H_0 : P_{xy} = P_x P_y$ (i.e., X and Y are independent) against the alternative $H_1 : P_{xy} \neq P_x P_y$. One way to test the null hypothesis with mean embedding is by comparing the RKHS distance between the embeddings of P_{xy} and $P_x P_y$. This requires a kernel $k\colon (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$. Assume that the kernel k on the joint domain is given by the product of kernels on the marginals. Then, the population quantity of the test statistic is $\mathrm{MMD}^2(P_{xy}, P_x P_y)$, which is known as the Hilbert-Schmidt Independence Criterion (HSIC) [Gretton et al., 2005b]. As in the case of the MMD two-sample test, an empirical estimator can be computed based on two kernels (one for X and one for Y), and its asymptotic distributions can be derived. Independence testing and HSIC will be discussed in details in Chapter 4.

Distribution Regression In distribution regression, we are given a paired sample $\{(P_i, y_i)\}_{i=1}^n$ drawn from a meta distribution $\mathcal{M}$, where P_i is a distribution defined on $\mathcal{X}$, and $y_i \in \mathbb{R}$ is the associated target output, the goal is to learn a map $P \mapsto y$ [Poczos et al., 2013]. One of many applications of distribution regression is, for instance, in predicting blood pressure (y_i) from a set of periodically measured health indicators (assumed to be represented by a distribution P_i). A flexible approach for distribution regression is by kernel ridge regression. In kernel ridge regression, regression can be performed provided that a kernel on the inputs can be defined. In the case of distribution regression, we require a kernel on distributions. Mean embedding offers a nonparametric and uniform way of defining a kernel on distributions. Let P and Q be two distributions, and $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel associated with RKHS $\mathcal{F}$. The

set kernel

$$\kappa(P, Q) = \langle \mu_P, \mu_Q \rangle_{\mathcal{F}} = \mathbb{E}_{\mathbf{x}\sim P}\mathbb{E}_{\mathbf{y}\sim Q} k(\mathbf{x}, \mathbf{y})$$

is one of the simplest kernels on distributions that can be defined with mean embeddings. The set kernel is linear in each of the two mean embeddings, analogous to the linear kernel $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$ for Euclidean input vectors. A natural extension of the Gaussian kernel in (2.3) to distributions is

$$\kappa_g(P, Q) = \exp\left(-\frac{\|\mu_P - \mu_Q\|_{\mathcal{F}}^2}{2\sigma^2}\right),$$

which is a Gaussian kernel on mean embeddings [Christmann and Steinwart, 2010]. Other nonlinear kernels on distributions can be found in Szabó et al. [2016, Table 1]. An advantage of defining kernels based on mean embeddings is that it is invariant to reparameterization of the input distributions. In Chapter 6, we will see an application of Gaussian process regression for distribution regression for predicting expectation propagation messages.

2.5 Properties of Kernels

We have seen in Section 2.1 that a kernel k directly characterizes the Hilbert space $\mathcal{F}$ of real-valued functions. In fact, functions in $\mathcal{F}(k)$ also inherit properties of k. We start with the boundedness.

Lemma 2.7 (Boundedness of kernels [Steinwart and Christmann, 2008, Lemma 4.23])**.** *Let $\mathcal{X}$ be a set and $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel with RKHS $\mathcal{F}$. Then, k is bounded if and only if $\|f\|_\infty = \sup_{\mathbf{x}\in\mathcal{X}} |f(\mathbf{x})| < \infty$ for all $f \in \mathcal{F}$.*

Lemma 2.7 states that the boundedness of the kernel k implies the boundedness of the functions in $\mathcal{F}(k)$, and vice versa. Continuity of the functions in the RKHS is also characterized by the continuity of the kernel, as stated in Lemma 2.8.

Lemma 2.8 (Continuity of kernels [Steinwart and Christmann, 2008, Lemma 4.28])**.** *Let $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel associated with the RKHS $\mathcal{F}$ on a topological space $\mathcal{X}$. Then, every $f \in \mathcal{F}$ is bounded and continuous if and only if k is bounded, and $\mathbf{v} \mapsto k(\mathbf{x}, \mathbf{v})$ is continuous for all $\mathbf{x} \in \mathcal{X}$.*

The "size" of an RKHS is also an important property for learning on distributions. In particular, if the function class defining an integral probability metric (IPM, see (2.8)) is large enough, then the IPM is a metric (rather than just a pseudometric). A useful class of large RKHSs is one which can approximate a continuous function up to any arbitrary accuracy. A kernel defining such an RKHS is called *universal* (Definition 2.9). We write $C(\mathcal{X})$ to denote the space of continuous functions, and write $C_b(\mathcal{X})$ to denote the space of bounded, continuous functions endowed with the uniform norm.

Definition 2.9 (Universal kernels [Steinwart and Christmann, 2008, Definition 4.52])**.** A continuous kernel k on a compact metric space $\mathcal{X}$ is called universal (also known as c-universal [Sriperumbudur et al., 2011]) if the associated RKHS $\mathcal{F}$ is dense in $C(\mathcal{X})$ i.e., for any function $g \in C(\mathcal{X})$ and any $\epsilon > 0$, there exists an $f \in \mathcal{F}$ such that $\|f - g\|_\infty \leq \epsilon$.

Gretton et al. [2012a, Theorem 5] showed that if k is universal on a compact metric space $\mathcal{X}$, then it is characteristic in the sense of Definition 2.6. The proof relies on the facts that an IPM ((2.8)) defined with the function class $C_b(\mathcal{X})$ is a metric on the space of Borel probability measures [Dudley, 2002, Lemma 9.3.2], and that any function from $C(\mathcal{X}) \supset C_b(\mathcal{X})$ can be well approximated by some function in a universal RKHS (by definition). When $\mathcal{X} \subset \mathbb{R}^d$ is compact, Steinwart and Christmann [2008, Corollary 4.58] shows that the exponential kernel $k(\mathbf{x}, \mathbf{y}) = \exp\left(\mathbf{x}^\top \mathbf{y}\right)$, and the Gaussian kernel in (2.3) are universal.

The compactness of $\mathcal{X}$ required in the definition of universal kernels can be restrictive and exclude many interesting spaces including $\mathbb{R}^d$. A variant of the c-universality is c_0-universality which does not require the domain to be a compact space. Before we define c_0-universality, we will need a few more definitions. Let $C_0(\mathcal{X})$ be the class of all continuous real-valued functions on $\mathcal{X}$ which vanish at infinity. Precisely, for any $\epsilon > 0$, and any $f \in C_0(\mathcal{X})$, the set $\{\mathbf{x} \in \mathcal{X} : |f(\mathbf{x})| \geq \epsilon\}$ is compact. When $\mathcal{X}$ is a normed vector space, the condition is equivalent to having $f(\mathbf{x}) \to 0$ as $\|\mathbf{x}\| \to \infty$ for all $f \in C_0(\mathcal{X})$. A space $\mathcal{X}$ is said to be locally compact if every point in $\mathcal{X}$ has a compact neighbourhood. A space $\mathcal{X}$ is Hausdorff if for any $\mathbf{x} \neq \mathbf{y} \in \mathcal{X}$, there exist a neighbourhood U of $\mathbf{x}$, and a neighbourhood V of $\mathbf{y}$ such that U and V are disjoint. Any metric space is Hausdorff. An example of a locally compact Hausdorff (LCH) space is $\mathbb{R}^d$ for $d \in \mathbb{N}$. We are ready to define c_0-universal kernels.

Definition 2.10 (c_0-kernels and c_0-universal kernels (Sriperumbudur et al. [2011, p. 2392], Carmeli et al. [2010, Definition 4.1]))**.** Let $\mathcal{X}$ be a locally compact Hausdorff (LCH) space. A kernel $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be a c_0-kernel if it is bounded with $k(\mathbf{x}, \cdot) \in C_0(\mathcal{X})$ for all $\mathbf{x} \in \mathcal{X}$. A kernel $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be c_0-universal if it is a c_0-kernel, and the RKHS $\mathcal{F}(k)$ is dense in $C_0(\mathcal{X})$ with respect to the uniform norm.

There is less restriction on the domain of a c_0-universal kernel compared with a c-universal kernel. c_0-universal kernels, however, need to vanish at infinity. Examples of c_0-universal kernels on $\mathbb{R}^d$ include the Gaussian (in (2.3)), Laplacian, B_{2l+1}-spline, inverse multiquadric, and the Matérn class. When $\mathcal{X}$ is compact, the c- and c_0-universality are equivalent. A c_0-universal kernel on an LCH space is characteristic [Sriperumbudur et al., 2011, p. 2398]. A c_0-kernel that is characteristic needs not be c_0-universal. However, this statement is true if the kernel is also translation invariant[3] on $\mathbb{R}^d$ [Sriperumbudur et al., 2011, p. 2397]. That is, a translation invariant,

[3] A kernel $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ on a vector space $\mathcal{X}$ is said to be translation invariant if there exists a function $\tilde{k} \colon \mathcal{X} \to \mathbb{R}$ such that $\tilde{k}(\mathbf{x} - \mathbf{y}) = k(\mathbf{x}, \mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. In words, the kernel k depends on only the difference of the two arguments.

characteristic c_0-kernel on $\mathbb{R}^d$ is c_0-universal. Translation invariant, c_0-kernels on $\mathbb{R}^d$ will be important in our discussion of independence testing in Chapter 4.

More recently, real analytic kernels were used to construct fast two-sample tests [Chwialkowski et al., 2015]. Real analytic kernels are defined in Definition 2.11.

Definition 2.11 (Real analytic kernels [Chwialkowski et al., 2015, p. 5])**.** Let $\mathcal{X} \subseteq \mathbb{R}^d$ be an open set. A kernel $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be real analytic (or simply analytic) if for all $\mathbf{v} \in \mathcal{X}$, $\mathbf{x} \mapsto k(\mathbf{x}, \mathbf{v})$ is a real analytic function on $\mathcal{X}$.

A real analytic kernel is necessarily infinitely differentiable, meaning that it is very smooth. An example of a real analytic kernel is the Gaussian kernel on an open set $\mathcal{X} \subseteq \mathbb{R}^d$. One useful consequence on the RKHS $\mathcal{F}(k)$ with a bounded, real analytic kernel k is that all functions in $\mathcal{F}(k)$ are real analytic (Lemma 2.12).

Lemma 2.12 (Analytic functions in RKHSs [Chwialkowski et al., 2015, Lemma 1][4])**.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be an open set. If a kernel $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is bounded and real analytic, then all functions in the RKHS associated with k are real analytic.*

A well-known property of a real analytic non-zero function is that its set of roots has measure zero (Lemma 2.13).

Lemma 2.13 (Roots of non-zero analytic functions have measure zero [Mityagin, 2015])**.** *Let f be a real analytic function on an open set $\mathcal{X} \subseteq \mathbb{R}^d$. If $f \neq 0$ (the zero function), then $\{\mathbf{x} \mid f(\mathbf{x}) = 0\}$ has zero Lebesgue measure.*

Since many kernel-based statistic tests boil down to determining whether an (empirically computed) RKHS function is a zero function, it turns out that Lemma 2.13 provides a fast alternative to computing the RKHS norm on the function. In Chapter 3, we will discuss an idea to exploit this fact to construct efficient linear-time tests.

[4] Lemma 1 in Chwialkowski et al. [2015] states this result only for when $\mathcal{X} = \mathbb{R}^d$. However, the same proof goes through with an open set $\mathcal{X} \subseteq \mathbb{R}^d$.

Chapter 3

Informative Features for Distinguishing Distributions

Summary We study two semimetrics on probability distributions [Chwialkowski et al., 2015], given as the sum of differences of expectations of analytic functions evaluated at spatial or frequency locations (i.e., features). The goal is to learn informative features for distinguishing two distributions. The features are learned so as to maximize the distinguishability of the distributions, by optimizing a lower bound on test power for a statistical test using these features. The result is a parsimonious and interpretable indication of how and where two distributions differ locally. We show that the empirical estimate of the test power criterion converges with increasing sample size, ensuring the quality of the returned features. In real-world benchmarks on high-dimensional text and image data, linear-time tests using the proposed semimetrics achieve comparable performance to the state-of-the-art quadratic-time maximum mean discrepancy test, while returning human-interpretable features that explain the test results.

3.1 Introduction

We address the problem of discovering features of distinct probability distributions, with which they can most easily be distinguished. The distributions may be in high dimensions, can differ in non-trivial ways (i.e., not simply in their means), and are observed only through i.i.d. samples. One application for such divergence measures is to model criticism, where samples from a trained model are compared with a validation sample: in the univariate case, through the KL divergence [Cinzia Carota and Polson, 1996], or in the multivariate case, by use of the maximum mean discrepancy (MMD) [Lloyd and Ghahramani, 2015] (see Section 2.3 for an introduction to MMD). In the latter work, the model output of the Automated Statistician [Lloyd et al., 2014] is compared with the original sample via a smooth witness function, which has largest amplitude where the sample probability mass differs most from the model. An alternative, interpretable analysis of a multivariate difference in distributions may be obtained by projecting onto a discriminative direction, such that the Wasserstein

distance on this projection is maximized [Mueller and Jaakkola, 2015]. Note that both recent works require low dimensionality, either explicitly (in the case of Lloyd and Gharamani, the function becomes difficult to plot in more than two dimensions), or implicitly in the case of Mueller and Jaakkola, in that a large difference in distributions must occur in projection along a particular one-dimensional axis. Distances between distributions in high dimensions may be more subtle, however, and it is of interest to find interpretable, distinguishing features of these distributions. For example, when divergence measures are used in adversarial learning for deep belief networks [Dziugaite et al., 2015, Li et al., 2015], it might be of interest to reveal explicitly the manner in which the distributions returned by samples from the model differ from the validation sample, which is challenging given the high dimensional outputs of the models.

In this chapter, we take a hypothesis testing approach to discovering features which best distinguish two multivariate probability measures P and Q, as observed by samples $\mathsf{X} := \{\mathbf{x}_i\}_{i=1}^n$ drawn independently and identically (i.i.d.) from P, and $\mathsf{Y} := \{\mathbf{y}_i\}_{i=1}^n \subset \mathcal{X} \subseteq \mathbb{R}^d$ from Q. Non-parametric two-sample tests based on RKHS distances [Moulines et al., 2008, Fromont et al., 2012, Gretton et al., 2012a] or energy distances [Székely and Rizzo, 2004, Baringhaus and Franz, 2004] have as their test statistic an integral probability metric, the Maximum Mean Discrepancy [Gretton et al., 2012a, Sejdinovic et al., 2013]. For this metric, a smooth witness function is computed, such that the amplitude is largest where the probability mass differs most [e.g. Gretton et al., 2012a, Figure 1]. Lloyd and Ghahramani [2015] used this witness function to compare the model output of the Automated Statistician [Lloyd et al., 2014] with a reference sample, yielding a visual indication of where the model fails. In high dimensions, however, the witness function cannot be plotted, and is less helpful. Furthermore, the witness function does not give an easily interpretable result for distributions with local differences in their characteristic functions. A more subtle shortcoming is that it does not provide a direct indication of the distribution features which, when compared, would maximize test power. Rather, it is the witness function *norm*, and (broadly speaking) its *variance* under the null, that determine test power.

Contributions Our approach builds on the analytic representations of probability distributions of Chwialkowski et al. [2015], where differences in expectations of analytic functions at particular spatial or frequency locations are used to construct a two-sample test statistic, which can be computed in linear time. Chwialkowski et al. [2015] showed that, despite the differences in these analytic functions being evaluated at random locations, the analytic tests have greater power than linear time tests based on subsampled estimates (see (2.14)) of the MMD [Gretton et al., 2012b, Zaremba et al., 2013]. Our first theoretical contribution, in Section 3.4.1, is to derive a lower bound on the test power, which can be maximized over the choice of test locations (features). We propose two novel variants, both of which significantly outperform the random feature choice of Chwialkowski et al.. The Mean Embedding (ME) test evaluates the

difference of mean embeddings at locations chosen to maximize the test power lower bound (i.e., spatial features); unlike the maxima of the MMD witness function, these features are directly chosen to maximize the distinguishability of the distributions, and take variance into account. The Smooth Characteristic Function (SCF) test uses as its statistic the difference of the two smoothed empirical characteristic functions, evaluated at points in the frequency domain so as to maximize the same criterion (i.e., frequency features). Optimization of the mean embedding kernels/frequency smoothing functions themselves is achieved on a held-out data set with the same consistent objective.

As our second theoretical contribution in Section 3.4.2, we prove that the empirical estimate of the test power criterion asymptotically converges to its population quantity uniformly over the class of Gaussian kernels, at the rate of $\mathcal{O}_p(n^{-1/4})$, where n is the sample size. Two important consequences follow: first, in testing, we obtain a more powerful test with fewer features. Second, we obtain a parsimonious and interpretable set of features that best distinguish the probability distributions. In Section 3.5, we provide experiments demonstrating that the proposed linear-time tests greatly outperform all considered linear time tests, and achieve performance that compares to or exceeds the more expensive quadratic-time MMD test [Gretton et al., 2012a]. Moreover, the new tests discover features of text data (NIPS proceedings) and image data (distinct facial expressions) which have a clear human interpretation, thus validating our feature elicitation procedure in these challenging high-dimensional testing scenarios.

3.2 Mean Embedding (ME) Test

Our approach of discovering distinguishing features is formulated as a nonparametric two-sample test based on the mean embedding (ME) and the smooth characteristic function (SCF) tests [Chwialkowski et al., 2015]. In this section, we review the ME test. The SCF test will be described in Section 3.3. Given two i.i.d. samples $\mathsf{X} := \{\mathbf{x}_i\}_{i=1}^n, \mathsf{Y} := \{\mathbf{y}_i\}_{i=1}^n$ from P and Q on $\mathcal{X} \subseteq \mathbb{R}^d$, respectively, the goal of a two-sample test is to decide whether P is different from Q on the basis of the samples. The task is formulated as a statistical hypothesis test proposing a null hypothesis $H_0 : P = Q$ (samples are drawn from the same distribution) against an alternative hypothesis $H_1 : P \neq Q$ (the sample generating distributions are different). A test calculates a test statistic $\hat{\lambda}_n$ from X and Y, and rejects H_0 if $\hat{\lambda}_n$ exceeds a predetermined test threshold (critical value). The threshold T_α is given by the $(1-\alpha)$-quantile of the distribution of $\hat{\lambda}_n$ under H_0 i.e., the null distribution, and α is the significance level of the test.

3.2.1 Unnormalized ME Statistic

The (unnormalized) ME test, in its simplest form, relies on a (random) metric on the space of Borel probability measures. Given a bounded, characteristic, integrable and real analytic kernel $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (see Section 2.5: Properties of Kernels),

consider the mean embeddings of P and Q given respectively by $\mu_P := \mathbb{E}_{\mathbf{x}\sim P} k(\mathbf{x}, \cdot)$ and $\mu_Q := \mathbb{E}_{\mathbf{y}\sim Q} k(\mathbf{y}, \cdot)$. The population unnormalized ME (UME) statistic is defined such that its square is

$$\mathrm{UME}^2(P, Q) = \frac{1}{J} \sum_{j=1}^{J} \left[\mu_P(\mathbf{v}_j) - \mu_Q(\mathbf{v}_j)\right]^2, \tag{3.1}$$

where $\mathcal{V} := \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$ is the set of *test locations* at which the witness function (i.e., difference of the two mean embeddings. See (2.9)) is evaluated. Chwialkowski et al. [2015, Theorem 2] shows that if $\mathcal{V}$ is drawn from a distribution with a density, then (3.1) is a *random metric* on the space of Borel probability measures. A random metric [Chwialkowski et al., 2015, Definition 1] simply means that it is a metric in the usual sense with qualification "almost-surely" attached to each property of the metric. More precisely, if $\mathcal{V}$ is drawn from a distribution η which has a density and whose support is a subset of $\mathcal{X}$, then η-almost surely $\mathrm{UME}(P, Q)$ is a metric. In particular, η-almost surely $\mathrm{UME}(P, Q) = 0$ if and only if $P = Q$. The η-almost-sureness means that there exists at least one setting of $\mathcal{V}$ such that $\mathrm{UME}(P, Q) = 0$ does not imply $P = Q$. However, if $\mathcal{V} \sim \eta$, then such an "unlucky" event will not happen (more precisely, the probability of such event is 0).

A consistent estimator of (3.1) is given by

$$\widehat{\mathrm{UME}^2} = \frac{1}{J} \sum_{j=1}^{J} \left[\hat{\mu}_P(\mathbf{v}_j) - \hat{\mu}_Q(\mathbf{v}_j)\right]^2 = \frac{1}{J} \sum_{j=1}^{J} \left[\bar{z}_{n,j}\right]^2 = \frac{1}{J} \bar{\mathbf{z}}_n^\top \bar{\mathbf{z}}_n, \tag{3.2}$$

where $\hat{\mu}_P(\mathbf{v}) := \frac{1}{n} \sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v})$ and $\hat{\mu}_Q(\mathbf{v}) := \frac{1}{n} \sum_{i=1}^{n} k(\mathbf{y}_i, \mathbf{v})$ are empirical mean embeddings of P and Q, respectively, $\bar{\mathbf{z}}_n := \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i \in \mathbb{R}^J$ and

$$\mathbf{z}_i := [k(\mathbf{x}_i, \mathbf{v}_1) - k(\mathbf{y}_i, \mathbf{v}_1), \ldots, k(\mathbf{x}_i, \mathbf{v}_J) - k(\mathbf{y}_i, \mathbf{v}_J)] \in \mathbb{R}^J.$$

If we assume that evaluation of $k(\mathbf{x}, \mathbf{v})$ costs $\mathcal{O}(d)$, then clearly (3.2) can be computed in $\mathcal{O}(dJn)$ time, which is linear in the sample size. We note that the pairing of $\mathbf{x}_i$ and $\mathbf{y}_i$ in $\mathbf{z}_i$ does not suggest any joint dependency between $\mathbf{x}_i$ and $\mathbf{y}_i$. Any arbitrary pairing yields an equivalent estimator. In principle, $\widehat{\mathrm{UME}^2}$ can be used as a test statistic for the two-sample test. However, under H_0, as $n \to \infty$, $\sqrt{n}\widehat{\mathrm{UME}^2}$ converges to a finite sum of weighted chi-squared variables that are dependent.[1] This distribution does not have a closed-form expression. Thus, determining $(1 - \alpha)$-quantile for the test threshold has to rely on simulations from the asymptotic null distribution, or a permutation test, both of which can be costly.

[1] Under H_0, for a fixed $\mathcal{V}$, by the central limit theorem, $\sqrt{n}\bar{z}_{n,j} \xrightarrow{d} \mathcal{N}(0, \mathbb{V}[k(\mathbf{x}, \mathbf{v}_j) - k(\mathbf{x}, \mathbf{v}_j)])$ for all $j \in \{1, \ldots, J\}$. Thus, $n\bar{z}_{n,j}^2 \xrightarrow{d} \mathbb{V}[k(\mathbf{x}, \mathbf{v}_j) - k(\mathbf{x}, \mathbf{v}_j)]\chi^2(1)$ for all $j \in \{1, \ldots J\}$. The variables $n\bar{z}_{n,1}^2, \ldots, n\bar{z}_{n,J}^2$ are dependent since they all depend on the samples X and Y.

3.2.2 Normalized ME (NME) Statistic

Intractability of the asymptotic null distribution was the motivation to consider the normalized ME statistic [Chwialkowski et al., 2015, Eq. 13]. The empirical normalized ME statistic (NME) is defined as

$$\widehat{\mathrm{NME}^2}(P, Q) = \hat{\lambda}_n := n\bar{\mathbf{z}}_n^\top \mathbf{S}_n^{-1}\bar{\mathbf{z}}_n, \tag{3.3}$$

where $\mathbf{S}_n := \frac{1}{n}\sum_{i=1}^n (\mathbf{z}_i - \bar{\mathbf{z}}_n)(\mathbf{z}_i - \bar{\mathbf{z}}_n)^\top \in \mathbb{R}^{J\times J}$ and $\bar{\mathbf{z}}_n$ is as in (3.2). The NME is a form of Hotelling's T-squared statistic [Anderson, 2003, Chapter 5]. The presence of the inverse covariance matrix $\mathbf{S}_n^{-1}$ is to decorrelate $\bar{z}_{n,1}, \ldots, \bar{z}_{n,J}$, so that the asymptotic null distribution is tractable. Asymptotic behaviors of $\hat{\lambda}_n$ are summarized in Proposition 3.1.

Proposition 3.1 (Asymptotic behaviors of $\hat{\lambda}_n$ [Chwialkowski et al., 2015, Proposition 2]). *Suppose* $\mathrm{UME}^2(P, Q) = 0$. *Then, for fixed* d, *as* $n \to \infty$, $\hat{\lambda}_n \xrightarrow{d} \chi^2(J)$, *the chi-squared random variable with* J *degrees of freedom. If* $\mathrm{UME}^2(P, Q) > 0$, *then for any fixed* r, $\mathbb{P}(\hat{\lambda}_n > r) \to 1$ *as* $n \to \infty$.

Proposition 3.1 states that the asymptotic null distribution of $\hat{\lambda}_n$ is $\chi^2(J)$ which is very simple. This asymptotic null distribution holds true regardless of P and Q. This implies that when n is sufficiently large, one can simply compute the $(1-\alpha)$-quantile of $\chi^2(J)$ for the test threshold. Under H_1, the probability of correctly rejecting H_0 approaches 1 as $n \to \infty$. These two facts mean that the NME test is consistent.

3.3 Smooth Characteristic Function (SCF) Test

The SCF test relies on the difference of smoothed (by a kernel) characteristic functions of P and Q on $\mathcal{X} \subseteq \mathbb{R}^d$, evaluated at J frequencies. Let $\varphi_P(\mathbf{t}) := \mathbb{E}_{\mathbf{x}\sim P}[e^{i\mathbf{t}^\top \mathbf{x}}]$ be the characteristic function of P where $i = \sqrt{-1}$. Similarly, $\varphi_Q(\mathbf{t}) := \mathbb{E}_{\mathbf{y}\sim P}[e^{i\mathbf{t}^\top \mathbf{y}}]$. To motivate the importance of the smoothing operation on the characteristic functions, consider the difference of (non-smoothed) characteristic functions [Epps and Singleton, 1986]:

$$\rho_\varphi^2(P, Q) := \frac{1}{J}\sum_{j=1}^J |\varphi_P(\mathbf{v}_j) - \varphi_Q(\mathbf{v}_j)|^2, \tag{3.4}$$

where $\mathcal{V} := \{\mathbf{v}_j\}_{j=1}^J$ is the set of frequency values at which the characteristic function difference is evaluated. The set $\mathcal{V}$ plays the same role (in the frequency domain) as the spatial test locations of the ME test. While (3.4) can be estimated efficiently in $\mathcal{O}(dJn)$ time (assuming the use of empirical characteristic functions), it turns out that $\rho_\varphi(P, Q)$ is only a pseudometric. In particular, $\rho_\varphi^2(P, Q) = 0$ does not always imply $P = Q$, as summarized in Proposition 3.2.

Proposition 3.2 (Chwialkowski et al. [2015, Proposition 1]). *Let* $J \in \mathbb{N}$ *and let* $\{v_j\}_{j=1}^J$ *be a sequence of real-valued i.i.d. random variables drawn from a distribution* η *with a density.*

Then, for any ϵ, there exists an uncountable set $\mathcal{A}$ of probability measures on the real line such that for any $P, Q \in \mathcal{A}$, we have $\mathbb{P}(\rho_\varphi^2(P,Q) = 0) \geq 1 - \epsilon$.

Proposition 3.2 implies that there are infinitely many distinct P, Q's that cannot be distinguished by $\rho_\varphi^2(P,Q)$. The SCF test remedies this by considering smooth characteristic functions instead (Definition 3.3).

Definition 3.3 (Smooth characteristic function [Chwialkowski et al., 2015, Definition 2])**.** A smooth characteristic function $\phi_P(\mathbf{v})$ of a distribution P is its characteristic function convolved with a real analytic, translation invariant, positive definite kernel $l\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (see also Section 2.5: Properties of Kernels):

$$\phi_P(\mathbf{v}) := \int_{\mathbb{R}^d} \varphi_P(\mathbf{w}) l(\mathbf{v} - \mathbf{w}) \, \mathrm{d}\mathbf{w}. \tag{3.5}$$

The population unnormalized test statistic of the SCF test is defined as

$$\mathrm{USCF}^2(P,Q) := \frac{1}{J} \sum_{j=1}^{J} |\phi_P(\mathbf{v}_j) - \phi_Q(\mathbf{v}_j)|^2. \tag{3.6}$$

In contrast to (3.4), the use of smooth characteristic functions in (3.6) greatly increases the class of distributions that can be distinguished.

Proposition 3.4 (USCF is a random metric [Chwialkowski et al., 2015, Theorem 1])**.** *Let $l\colon \mathcal{X} \to \mathbb{R}$ be an integrable, real analytic, translation invariant kernel (i.e., $(\mathbf{x}, \mathbf{y}) \mapsto l(\mathbf{x} - \mathbf{y})$ defines a positive definite kernel on $\mathcal{X} \times \mathcal{X}$), whose inverse Fourier transform is strictly greater than zero. Let $\{\mathbf{v}_j\}_{j=1}^J$ be realizations from a distribution η that has a density. Then, for any $J > 0$, η-almost surely, $\mathrm{USCF}^2(P,Q)$ is a metric on the space of probability measures that have integrable characteristic functions.*

Proposition 3.4 guarantees that for any distributions P, Q that have integrable characteristic functions, almost surely $\mathrm{USCF}^2(P,Q) = 0$ if and only if $P = Q$. Thus, an empirical estimate of (3.6) can be used as a test statistic for two-sample testing. The smooth characteristic function in (3.5) has an equivalent expression [Chwialkowski et al., 2015, Proposition 3] which avoids computationally difficult convolution:

$$\phi_P(\mathbf{v}) := \mathbb{E}_{\mathbf{x} \sim P}[\exp(i\mathbf{v}^\top \mathbf{x}) \hat{l}(\mathbf{x})], \tag{3.7}$$

where $\hat{l}(\mathbf{x}) = \int_{\mathbb{R}^d} \exp(-i\mathbf{u}^\top \mathbf{x}) l(\mathbf{u}) \, \mathrm{d}\mathbf{u}$ is the inverse Fourier transform of $l(\mathbf{x})$. The plug-in estimator of (3.7) is straightforwardly given by $\hat{\phi}_P(\mathbf{v}) = \frac{1}{n} \sum_{i=1}^n \exp(i\mathbf{v}^\top \mathbf{x}_i) \hat{l}(\mathbf{x}_i)$. It follows that a consistent estimator of (3.6) is

$$\widehat{\mathrm{USCF}^2} = \frac{1}{J} \sum_{j=1}^{J} \left[\hat{\phi}_P(\mathbf{v}_j) - \hat{\phi}_Q(\mathbf{v}_j)\right]^2 = \frac{1}{J} \sum_{j=1}^{J} \left[\overline{z}_{n,j}\right]^2 = \frac{1}{J} \overline{\mathbf{z}}_n^\top \overline{\mathbf{z}}_n, \tag{3.8}$$

where

$$\mathbf{z}_i := [\hat{l}(\mathbf{x}_i)\sin(\mathbf{x}_i^\top \mathbf{v}_j) - \hat{l}(\mathbf{y}_i)\sin(\mathbf{y}_i^\top \mathbf{v}_j), \hat{l}(\mathbf{x}_i)\cos(\mathbf{x}_i^\top \mathbf{v}_j) - \hat{l}(\mathbf{y}_i)\cos(\mathbf{y}_i^\top \mathbf{v}_j)]_{j=1}^J \in \mathbb{R}^{2J},$$

and $\overline{\mathbf{z}}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i$. Here, we have stacked together the real and imaginary parts of $\exp(i\mathbf{v}^\top \mathbf{x})\hat{l}(\mathbf{x})$ and $\exp(i\mathbf{v}^\top \mathbf{y})\hat{l}(\mathbf{y})$ to avoid complex numbers. Clearly, (3.8) takes the same form as the unnormalized ME statistic in (3.2), where the change is the definition of $\mathbf{z}_i$. Note that the length of $\mathbf{z}_i$ in the case of SCF test is $2J$. As in the case of the normalized ME statistic in (3.3), we can consider the normalized SCF statistic (NSCF) to get a tractable asymptotic null distribution. The NSCF statistic takes the same form in the normalized ME statistic:

$$\widehat{\text{NSCF}^2}(P, Q) = \hat{\lambda}_n := n\overline{\mathbf{z}}_n^\top \mathbf{S}_n^{-1}\overline{\mathbf{z}}_n.$$

Here we intentionally use the same notation $\hat{\lambda}_n$ for both the normalized ME and SCF statistics, since they share many common properties. The distinction will be clear from the context. The asymptotic null distribution of $\widehat{\text{NSCF}^2}$ is $\chi^2(2J)$. The test based on $\widehat{\text{NSCF}^2}$ is consistent when P and Q have integrable characteristic functions [Chwialkowski et al., 2015, Proposition 4]. We will use J' to refer to the degrees of freedom of the chi-squared distribution i.e., $J' = J$ for the ME test, and $J' = 2J$ for the SCF test.

3.4 Proposal: Interpretable Two-Sample Tests

This section describes our contributions. We extend the ME and SCF tests to learn discriminative features $\mathcal{V} = \{\mathbf{v}_j\}_{j=1}^J$ by maximizing a proxy for the test power (Section 3.4.1). The learned features give a visual indication of where the two distributions differ. We then derive a simple empirical optimization objective, and theoretically justify its convergence (Section 3.4.2). In the followings, we will interchangeably use the terms test locations and features to refer to $\mathcal{V}$.

In our study, we modify the ME and SCF statistics by adding to $\mathbf{S}_n$ an identity matrix $\mathbf{I}$ scaled by a regularization parameter $\gamma_n > 0$, giving

$$\hat{\lambda}_n := n\overline{\mathbf{z}}_n^\top \left(\mathbf{S}_n + \gamma_n \mathbf{I}\right)^{-1} \overline{\mathbf{z}}_n, \tag{3.9}$$

for stability of the matrix inverse. Using multivariate Slutsky's theorem, under H_0, $\hat{\lambda}_n$ still asymptotically follows $\chi^2(J')$ provided that $\gamma_n \to 0$ as $n \to \infty$. Consistency of the two tests remains true. We start by describing the criterion we use to learn discriminative features.

3.4.1 A Test Power Lower Bound and Feature Learning

We propose optimizing the test locations $\mathcal{V}$ and kernel parameters (jointly referred to as θ) by maximizing a lower bound on the test power in Proposition 3.5. This criterion offers a simple objective function for fast parameter tuning. The bound may be of

independent interest in other Hotelling's T-squared statistics, since apart from the Gaussian case [e.g. Bilodeau and Brenner, 2008, Ch. 8], the characterization of such statistics under the alternative distribution is challenging. We use $\mathbb{E}_{\mathbf{xy}}$ as a shorthand for $\mathbb{E}_{\mathbf{x}\sim P}\mathbb{E}_{\mathbf{y}\sim Q}$ and let $\|\cdot\|_F$ be the Frobenius norm.

Proposition 3.5 (Lower bound on the test power)**.** *Define* $\boldsymbol{\mu} := \mathbb{E}_{\mathbf{xy}}[\mathbf{z}_1]$, $\boldsymbol{\Sigma} := \mathbb{E}_{\mathbf{xy}}[(\mathbf{z}_1 - \boldsymbol{\mu})(\mathbf{z}_1 - \boldsymbol{\mu})^\top]$, *and* $\lambda_n := n\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ *(the population counterpart of* $\hat{\lambda}_n$*). Let* $\mathbb{V}$ *be a collection in which each element is a set of* J *test locations.*

- *For the ME test, let* $\mathcal{K}$ *be a uniformly bounded (i.e., there exists* $B < \infty$ *such that* $\sup_{k\in\mathcal{K}} \sup_{(\mathbf{x},\mathbf{y})\in\mathcal{X}^2} |k(\mathbf{x},\mathbf{y})| \leq B$*) family of* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *measurable kernels.*
- *For the SCF test, let* $\mathcal{K}$ *be a class of translation-invariant kernels such that* $\mathcal{L} = \{\mathbf{x} \mapsto \hat{k}(\mathbf{x}) : k \in \mathcal{K}\}$ *is uniformly bounded (i.e.,* $\exists B < \infty$ *such that* $\sup_{f\in\mathcal{L}} \sup_{\mathbf{x}\in\mathcal{X}} |f(\mathbf{x})| \leq B$*) family, where* $\hat{k}(\mathbf{x}) := \int e^{-i\mathbf{w}^\top\mathbf{x}} k(\mathbf{w})\,\mathrm{d}\mathbf{w}$.

Assume that $\tilde{c} := \sup_{\mathcal{V}\in\mathbb{V},k\in\mathcal{K}} \|\boldsymbol{\Sigma}^{-1}\|_F < \infty$. *Then, for any* $\mathcal{V} \in \mathbb{V}$, *for large* n, *the test power* $\mathbb{P}\left(\hat{\lambda}_n \geq T_\alpha\right)$ *of both tests satisfies* $\mathbb{P}\left(\hat{\lambda}_n \geq T_\alpha\right) \geq L(\lambda_n)$ *where*

$$L(\lambda_n) := 1 - 2e^{-\frac{(\lambda_n - T_\alpha)^2}{3^2\cdot 8B^2\bar{c}_2^2 J' n}} - 2e^{-\frac{(\gamma_n(\lambda_n - T_\alpha)(n-1) - 24B^2\bar{c}_1 J' n)^2}{3^2\cdot 32B^4\bar{c}_1^2 J'^2 n(2n-1)^2}} - 2e^{-\frac{\left((\lambda_n - T_\alpha)/3 - \bar{c}_3 n\gamma_n\right)^2\gamma_n^2}{32B^4 J'^2\bar{c}_1^2 n}},$$

$\bar{c}_1 := 4B^2J'\sqrt{J}\tilde{c}, \bar{c}_2 := 4B\sqrt{J'}\tilde{c}$, *and* $\bar{c}_3 := 4B^2J'\tilde{c}^2$. *For the ME test,* $J' = J$. *For the SCF test,* $J' = 2J$. *For large* n, $L(\lambda_n)$ *is increasing in* λ_n.

Proof (sketch). The idea is to construct a bound for $|\hat{\lambda}_n - \lambda_n|$ which involves bounding $\|\bar{\mathbf{z}}_n - \boldsymbol{\mu}\|_2$ and $\|\mathbf{S}_n - \boldsymbol{\Sigma}\|_F$ separately using Hoeffding's inequality. The result follows after a reparameterization of the bound on $\mathbb{P}(|\hat{\lambda}_n - \lambda_n| \geq t)$ to have $\mathbb{P}\left(\hat{\lambda}_n \geq T_\alpha\right)$. See Section 3.B for details. □

Proposition 3.5 suggests that for large n it is sufficient to maximize λ_n to maximize the lower bound on the test power of both ME and SCF tests. Assume that k is characteristic [Sriperumbudur et al., 2011] for the ME test. It can be shown that $\lambda_n = 0$ if and only if $P = Q$ i.e., λ_n is a semimetric for P and Q. In this sense, one can see λ_n as encoding the ease of rejecting H_0. The higher λ_n, the easier for the test to correctly reject H_0 when H_1 holds. This observation justifies the use of λ_n as a maximization objective for parameter tuning.

Feature Learning The statistic $\hat{\lambda}_n$ for both ME and SCF tests depends on a set of test locations $\mathcal{V}$ and a kernel parameter σ. We propose setting

$$\theta := \{\mathcal{V}, \sigma\} = \arg\max_\theta \lambda_n = \arg\max_\theta \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}. \tag{3.10}$$

The optimization of θ brings two benefits: first, it significantly increases the probability of rejecting H_0 when H_1 holds; second, the learned test locations act as discriminative features allowing an interpretation of how the two distributions differ. To avoid creating a dependency between θ and the data used for testing (which would affect

the null distribution), we split the data into two disjoint sets. Let $\mathsf{D} := (\mathsf{X}, \mathsf{Y})$ and $\mathsf{D}^{tr}, \mathsf{D}^{te} \subset \mathsf{D}$ such that $\mathsf{D}^{tr} \cap \mathsf{D}^{te} = \emptyset$ and $\mathsf{D}^{tr} \cup \mathsf{D}^{te} = \mathsf{D}$. In practice, since $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown, for parameter tuning, we use $\hat{\lambda}^{tr}_{n/2}$ to estimate λ_n, where $\hat{\lambda}^{tr}_{n/2}$ is the test statistic computed on the training set D^{tr}. Specifically, the optimization in (3.10) is approximated by

$$\theta^{tr} \approx \arg\max_{\theta} \overline{\mathbf{z}}_n^{tr\top} (\mathbf{S}_n^{tr} + \gamma_n \mathbf{I})^{-1} \overline{\mathbf{z}}_n^{tr},$$

where $\overline{\mathbf{z}}_n^{tr}$ and $\mathbf{S}_n^{tr}$ are $\mathbf{z}_n$ and $\mathbf{S}_n$ estimated using the training sample D^{tr}, and θ^{tr} refers to θ optimized on the training data. The regularization nparameter γ_n is not optimized; we set γ_n to be as small as possible while being large enough to ensure that $(\mathbf{S}_n^{tr} + \gamma_n \mathbf{I})^{-1}$ can be stably computed. For simplicity, we assume that each of D^{tr} and D^{te} has half of the samples in D. We perform an optimization of θ^{tr} with gradient ascent algorithm on $\hat{\lambda}^{tr}_{n/2}(\theta)$. The actual two-sample test is performed using the test statistic $\hat{\lambda}^{te}_{n/2}(\theta^{tr})$ computed on D^{te}. The full algorithm for the two proposed tests from parameter tuning to the actual two-sample testing is given in Algorithm 3.1.

Algorithm 3.1 Optimizing parameters and testing in the ME and SCF tests

Require: Two samples X, Y, significance level α, and number J of test locations.

1: Split $\mathsf{D} := (\mathsf{X}, \mathsf{Y})$ into disjoint training and test sets, D^{tr} and D^{te}, of the same size n^{te}.
2: Optimize parameters $\theta^{tr} = \arg\max_{\theta} \hat{\lambda}^{tr}_{n/2}(\theta)$ where $\hat{\lambda}^{tr}_{n/2}(\theta)$ is computed with the training set D^{tr}.
3: Set T_α to the $(1-\alpha)$-quantile of $\chi^2(J')$.
4: Compute the test statistic $\hat{\lambda}^{te}_{n/2}(\theta^{tr})$ using D^{te}.
5: Reject H_0 if $\hat{\lambda}^{te}_{n/2}(\theta^{tr}) > T_\alpha$.

We note that optimizing parameters by maximizing a test power proxy [Gretton et al., 2012b] is valid under both H_0 and H_1 as long as the data used for parameter tuning and for testing are disjoint. If H_0 is true, then Proposition 3.1 guarantees that $\hat{\lambda}^{te}_{n/2}(\theta) \xrightarrow{d} \chi^2(J')$ for any θ which is independent of the test data. Splitting the data into two disjoint sets guarentees the independence. Thus, the optimized θ^{tr} does not change the null distribution. Also, the rejection threshold T_α depends on only J' and is independent of the optimized parameters. If, instead, the optimization is performed on the same data used for testing (i.e., consider $\hat{\lambda}^{te}_{n}(\theta^{te})$), then the asymptotic null distribution of $\chi^2(J')$ no longer holds. In this case, the asymptotic rate of false rejection of H_0 will be larger than α if the test threshold is still set to $(1-\alpha)$-quantile of $\chi^2(J')$.

Optimization Assume that the Gaussian kernel $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma^2}\right)$ is used. Then, the parameter tuning objective $\hat{\lambda}^{tr}_{n/2}(\theta)$ is a function of θ consisting of one positive real σ and J test locations, each having d dimensions. The parameters θ can thus be regarded as a $Jd+1$ Euclidean vector. We take the derivative of $\hat{\lambda}^{tr}_{n/2}(\theta)$ with respect to θ, and use gradient ascent to maximize it. Since σ^2 is always non-negative, positivity constraint on σ does not need to be explicitly enforced. J is pre-specified

and fixed. For the ME test, we initialize the test locations with realizations from two multivariate normal distributions fitted to samples from P and Q; this ensures that the initial locations are well supported by the data. For the SCF test, initialization using the standard normal distribution is found to be sufficient. We emphasize that both the optimization and testing are linear in n. Computing the regularized statistic in (3.9) for both ME and SCF tests costs $\mathcal{O}(J'^3 + J'^2 n + dJ'n)$, and the optimization costs $\mathcal{O}(J'^3 + dJ'^2 n)$ per gradient ascent iteration.

3.4.2 Convergence of the Normalized ME Test Power Criterion

Since we use an empirical estimate $\hat{\lambda}^{tr}_{n/2}$ in place of the population power criterion λ_n for parameter optimization, we give a finite-sample bound in Theorem 3.6 guaranteeing the convergence of $\bar{\mathbf{z}}_n^\top (\mathbf{S}_n + \gamma_n I)^{-1} \bar{\mathbf{z}}_n$ to $\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$ as n increases, in the case of the ME test. The convergence is uniform over all kernels $k \in \mathcal{K}$ (a family of uniformly bounded kernels) and all test locations in an appropriate class $\mathbb{V}$.

Theorem 3.6 (Convergence of the power criterion of the NME test)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a measurable set, and $\mathbb{V}$ be a collection in which each element is a set of J test locations. For a class of kernels $\mathcal{K}$ on $\mathcal{X} \subseteq \mathbb{R}^d$, define*

$$\mathcal{F}_1 := \{\mathbf{x} \mapsto k(\mathbf{x}, \mathbf{v}) \mid k \in \mathcal{K}, \mathbf{v} \in \mathcal{X}\}, \tag{3.11}$$
$$\mathcal{F}_2 := \{\mathbf{x} \mapsto k(\mathbf{x}, \mathbf{v}) k(\mathbf{x}, \mathbf{v}') \mid k \in \mathcal{K}, \mathbf{v}, \mathbf{v}' \in \mathcal{X}\},$$
$$\mathcal{F}_3 := \{(\mathbf{x}, \mathbf{y}) \mapsto k(\mathbf{x}, \mathbf{v}) k(\mathbf{y}, \mathbf{v}') \mid k \in \mathcal{K}, \mathbf{v}, \mathbf{v}' \in \mathcal{X}\}. \tag{3.12}$$

Assume

1. *$\mathcal{K}$ is a uniformly bounded (by B) family of $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ measurable kernels,*
2. *$\tilde{c} := \sup_{\mathcal{V} \in \mathbb{V}} \sup_{k \in \mathcal{K}} \|\boldsymbol{\Sigma}^{-1}\|_F < \infty$, and*
3. *$\mathcal{F}_i = \{f_{\theta_i} \mid \theta_i \in \Theta_i\}$ is VC-subgraph [van der Vaart and Wellner, 2000] with VC-index $VC(\mathcal{F}_i) < \infty$, and $\mathcal{F}_i$ is a separable Carathéodory family (i.e., Θ_i is in a separable metric space, and $\theta_i \mapsto f_{\theta_i}(\mathbf{x})$ is continuous for all $\mathbf{x} \in \mathcal{X}$), for all $i = 1, 2, 3$.*

Let $\bar{c}_1 := 4B^2 J \sqrt{J} \tilde{c}, \bar{c}_2 := 4B\sqrt{J}\tilde{c}$, and $\bar{c}_3 := 4B^2 J \tilde{c}^2$. Let C_i-s $(i = 1, 2, 3)$ be the universal constants associated to $\mathcal{F}_i$-s according to Theorem 2.6.7 in van der Vaart and Wellner [2000]. Then for any $\delta \in (0, 1)$ with probability at least $1 - \delta$,

$$\sup_{\mathcal{V} \in \mathbb{V}} \sup_{k \in \mathcal{K}} \left| \bar{\mathbf{z}}_n^\top (\mathbf{S}_n + \gamma_n \mathbf{I})^{-1} \bar{\mathbf{z}}_n - \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right|$$
$$\leq 2 T_{\mathcal{F}_1} \left(\frac{2}{\gamma_n} \bar{c}_1 B J \frac{2n-1}{n-1} + \bar{c}_2 \sqrt{J} \right) + \frac{2}{\gamma_n} \bar{c}_1 J (T_{\mathcal{F}_2} + T_{\mathcal{F}_3}) + \frac{8}{\gamma_n} \frac{\bar{c}_1 B^2 J}{n-1} + \bar{c}_3 \gamma_n,$$

where

$$T_{\mathcal{F}_j} = \frac{16\sqrt{2} B^{\zeta_j}}{\sqrt{n}} \left(2\sqrt{\log \left[C_j \times VC(\mathcal{F}_j)(16e)^{VC(\mathcal{F}_j)} \right]} + \frac{\sqrt{2\pi [VC(\mathcal{F}_j) - 1]}}{2} \right) + B^{\zeta_j} \sqrt{\frac{2 \log(5/\delta)}{n}},$$

for $j = 1, 2, 3$ *and* $\zeta_1 = 1, \zeta_2 = \zeta_3 = 2$.

Proof (sketch). The idea is to lower bound the difference with an expression involving $\sup_{\mathcal{V}\in\mathbb{V}} \sup_{k\in\mathcal{K}} \|\overline{\mathbf{z}}_n - \boldsymbol{\mu}\|_2$ and $\sup_{\mathcal{V}\in\mathbb{V}} \sup_{k\in\mathcal{K}} \|\mathbf{S}_n - \boldsymbol{\Sigma}\|_F$. These two quantities can be seen as suprema of empirical processes, and can be bounded by Rademacher complexities of their respective function classes (i.e., $\mathcal{F}_1, \mathcal{F}_2$, and $\mathcal{F}_3$). Finally, the Rademacher complexities can be upper bounded using Dudley entropy bound and VC subgraph properties of the function classes. Proof details are given in Section 3.A. □

Theorem 3.6 implies that if we set $\gamma_n = \mathcal{O}(n^{-1/4})$, then we have

$$\sup_{\mathcal{V}\in\mathbb{V}} \sup_{k\in\mathcal{K}} \left| \overline{\mathbf{z}}_n^\top (\mathbf{S}_n + \gamma_n \mathbf{I})^{-1} \overline{\mathbf{z}}_n - \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right| = \mathcal{O}_p(n^{-1/4})$$

as the rate of convergence. A kernel class $\mathcal{K}$ satisfying the three conditions of Theorem 3.6 is the widely used isotropic Gaussian kernel class

$$\mathcal{K}_g = \left\{ k_\sigma : (\mathbf{x}, \mathbf{y}) \mapsto \exp\left(-(2\sigma^2)^{-1} \|\mathbf{x} - \mathbf{y}\|^2\right) \mid \sigma \in [g_l, g_u] \right\}, \tag{3.13}$$

for any (g_l, g_u) such that $0 < g_l < g_u < \infty$. In fact, a generic isotropic Gaussian kernel class in which $\sigma > 0$ (i.e., not necessarily restricted to a compact set) satisfies conditions 1 and 3 of Theorem 3.6 as shown in Lemma 3.9. To further guarantee that $\tilde{c} = \sup_{\mathcal{V}\in\mathbb{V}} \sup_{k\in\mathcal{K}} \|\boldsymbol{\Sigma}^{-1}\|_F < \infty$ (condition 2 of Theorem 3.6), it is sufficient that the Gaussian width σ in $\mathcal{K}_g$ be constrained to be in a compact set. Also, for $\mathbb{V}$, consider $\mathbb{V}$ such that any two test locations in any $\mathcal{V} \in \mathbb{V}$ are at least ϵ distance apart (i.e., distinct test locations), and that all test locations have a norm uniformly bounded by ζ, for some $\epsilon, \zeta > 0$. Then, for any non-degenerate P, Q, we have $\tilde{c} < \infty$ since $(k, \mathcal{V}) \mapsto \lambda_n$ is continuous on $\mathcal{K} \times \mathbb{V}$, and thus attains its supremum over compact sets $\mathcal{K}$ and $\mathbb{V}$. The distinctiveness of $\{\mathbf{v}_j\}_{j=1}^J$ is a necessary condition for $\boldsymbol{\Sigma}$ to be invertible.

Briefly, in Lemma 3.9, we show that $\mathcal{K} = \mathcal{K}_g$ is uniformly bounded with $B = 1$ (condition 1 of Theorem 3.6), the induced classes $\mathcal{F}_1, \mathcal{F}_2$, and $\mathcal{F}_3$ are VC-subgraphs with VC-indices linear in d, and are such that $\theta_i \mapsto f_{\theta_i}$ is continuous. Specifically, in condition 3 of Theorem 3.6, $\Theta_1 = \{(k, \mathbf{v}) \mid k \in \mathcal{K}_g, \mathbf{v} \in \mathcal{X}\}$ and $\Theta_2 = \Theta_3 = \{(k, \mathbf{v}, \mathbf{v}') \mid k \in \mathcal{K}_g, \mathbf{v}, \mathbf{v}' \in \mathcal{X}\}$.

3.5 Experiments

In this section, we demonstrate the effectiveness of the proposed methods on both toy and real problems. We study seven kernel-based two-sample tests, all using the isotropic Gaussian kernel class $\mathcal{K}_g$ in (3.13). For the SCF test, we set $\hat{l}(\mathbf{x}) = k(\mathbf{x}, \mathbf{0})$ where $k \in \mathcal{K}_g$. Denote by ME-full and SCF-full the ME and SCF tests whose test locations $\mathcal{V}$ and the Gaussian width σ are fully optimized using gradient ascent on a separate training sample (D^{tr}) of the same size as the test set (D^{te}). ME-grid and SCF-grid are as in Chwialkowski et al. [2015] where only the Gaussian width

is optimized by a grid search,[2] and the test locations are randomly drawn from a multivariate normal distribution. Specifically, for ME-grid, the test locations are drawn from a multivariate normal distribution fitted to the training data. For SCF-grid, the test locations (in the frequency domain) are drawn from the standard multivariate normal distribution. MMD-quad (quadratic-time) and MMD-lin (linear-time) refer to the nonparametric tests based on maximum mean discrepancy of Gretton et al. [2012a] (see Section 2.3: Maximum Mean Discrepancy), where to ensure a fair comparison, the Gaussian kernel width is also chosen so as to maximize a criterion for the test power on training data [Gretton et al., 2012b, Sutherland et al., 2016]. For MMD-quad, since its null distribution is given by an infinite sum of weighted chi-squared variables (no closed-form quantiles), in each trial we randomly permute the two samples 400 times to approximate the null distribution [Gretton et al., 2012a]. Finally, T^2 is the standard two-sample Hotelling's T-squared test, which serves as a baseline with Gaussian assumptions on P and Q.

In all the following experiments, each problem is repeated for 500 trials. For toy problems, new samples are generated from the specified P, Q distributions in each trial. For real problems, samples are partitioned randomly into training and test sets in each trial. In all of the simulations, we report an estimate of the rejection rate i.e., the proportion of the number of times that $\hat{\lambda}^{te}_{n/2}$ is above T_α. This quantity is an estimate of type-I error under H_0, and corresponds to test power when H_1 is true. We set $\alpha = 0.01$ in all the experiments.

3.5.1 Informative Features: Simple Demonstration

We begin with a demonstration that the proxy $\hat{\lambda}^{tr}_{n/2}(\theta)$ for the test power is informative for revealing the difference of the two distributions in the ME test. We consider the Gaussian Mean Difference (GMD) problem (see Table 3.1), where both P and Q are two-dimensional normal distributions with the difference in means. We use $J = 2$ test locations $\mathbf{v}_1$ and $\mathbf{v}_2$, where $\mathbf{v}_1$ is fixed to the location indicated by the black triangle in Figure 3.1. The contour plot shows $\mathbf{v}_2 \mapsto \hat{\lambda}^{tr}_{n/2}(\mathbf{v}_1, \mathbf{v}_2)$.

Figure 3.1 (left) suggests that $\hat{\lambda}^{tr}_{n/2}$ is maximized when $\mathbf{v}_2$ is placed in either of the two regions that captures the difference of the two samples i.e., the region in which the probability masses of P and Q have less overlap. Figure 3.1 (right), we consider placing $\mathbf{v}_1$ in one of the two key regions. In this case, the contour plot shows that $\mathbf{v}_2$ should be placed in the other region to maximize $\hat{\lambda}^{tr}_{n/2}$, implying that placing multiple test locations in the same neighborhood will not increase the discriminability. The two modes on the left and right suggest two ways to place the test location in a region that reveals the difference. The non-convexity of the $\hat{\lambda}^{tr}_{n/2}$ is an indication of many informative ways to detect differences of P and Q, rather than a drawback. A convex objective would not capture this multimodality.

[2] Chwialkowski et al. [2015] chooses the Gaussian width that minimizes the median of the p-values, a heuristic that does not directly address test power. Here, we perform a grid search to choose the best Gaussian width by maximizing $\hat{\lambda}^{tr}_{n/2}$ as done in ME-full and SCF-full.

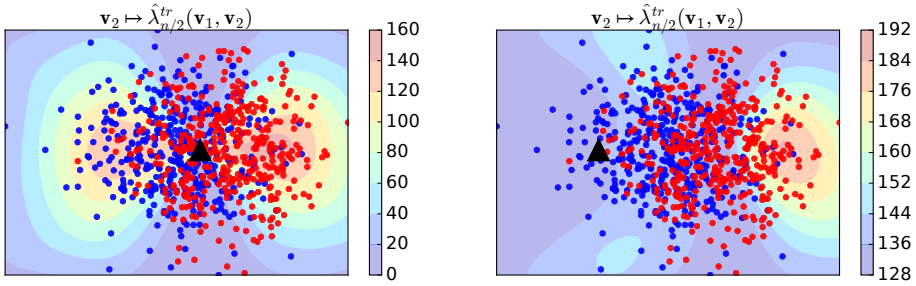
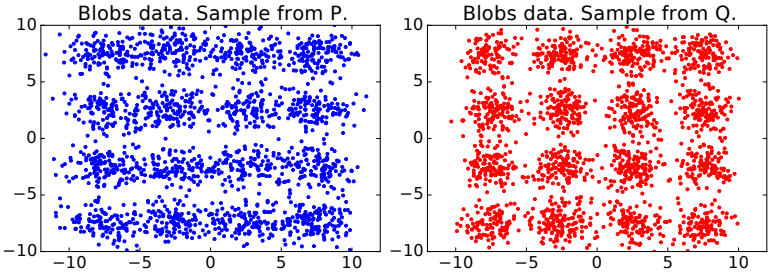


Figure 3.1: A contour plot of $\hat{\lambda}^{tr}_{n/2}$ as a function of $\mathbf{v}_2$ when $J = 2$ and $\mathbf{v}_1$ is fixed (black triangle). The objective $\hat{\lambda}^{tr}_{n/2}$ is high in the regions that reveal the difference of the two samples.

Table 3.1: Four toy problems. H_0 holds only in SG.

Data	P	Q
SG	$\mathcal{N}(\mathbf{0}_d, I_d)$	$\mathcal{N}(\mathbf{0}_d, I_d)$
GMD	$\mathcal{N}(\mathbf{0}_d, I_d)$	$\mathcal{N}((1, 0, \ldots, 0)^\top, I_d)$
GVD	$\mathcal{N}(\mathbf{0}_d, I_d)$	$\mathcal{N}(\mathbf{0}_d, \mathrm{diag}(2, 1, \ldots, 1))$
Blobs	Gaussian mixtures in $\mathbb{R}^2$ as studied in Chwialkowski et al. [2015], Gretton et al. [2012b].	

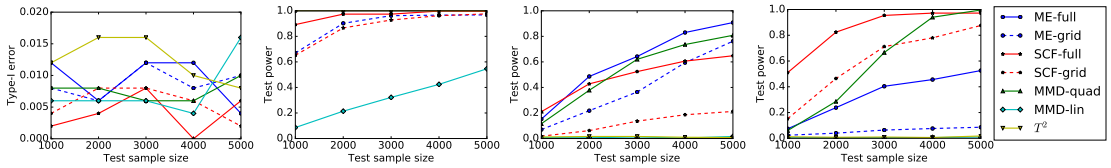


3.5.2 Test Power Vs. Sample Size n

We now demonstrate the rate of increase of test power with sample size. When the null hypothesis holds, the type-I error stays at the specified level α. We consider the following four toy problems: Same Gaussian (SG), Gaussian mean difference (GMD), Gaussian variance difference (GVD), and Blobs. The specifications of P and Q are summarized in Table 3.1. In the Blobs problem, P and Q are defined as a mixture of Gaussian distributions arranged on a 4×4 grid in $\mathbb{R}^2$. This problem is challenging as the difference of P and Q is encoded at a much smaller length scale compared to the global structure [Gretton et al., 2012b]. Specifically, the eigenvalue ratio for the covariance of each Gaussian distribution is 2.0 in P, and 1.0 in Q. We set $J = 5$.

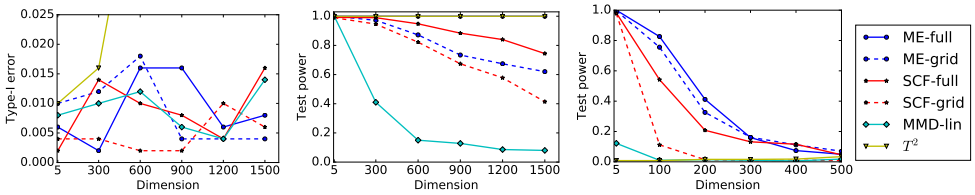
The results are shown in Figure 3.2 where type-I error (for SG problem), and test power (for GMD, GVD and Blobs problems) are plotted against test sample size. A number of observations are worth noting. In the SG problem, we see that the type-I error roughly stays at the specified level: the rate of rejection of H_0 when it is true is roughly at the specified level $\alpha = 0.01$.

GMD with 100 dimensions turns out to be an easy problem for all the tests except MMD-lin. In the GVD and Blobs cases, ME-full and SCF-full achieve substantially



(a) SG. $d = 50$. (b) GMD. $d = 100$. (c) GVD. $d = 50$. (d) Blobs.

Figure 3.2: Plots of type-I error/test power against the test sample size n^{te} in the four toy problems.



(a) SG (b) GMD (c) GVD

Figure 3.3: Plots of type-I error/test power against the dimensions d in the four toy problems in Table 3.1.

higher test power than ME-grid and SCF-grid, respectively, suggesting a clear advantage from optimizing the test locations. Remarkably, ME-full consistently outperforms the quadratic-time MMD across all test sample sizes in the GVD case. When the difference of P and Q is subtle as in the Blobs problem, ME-grid, which uses randomly drawn test locations, can perform poorly (see Figure 3.2d) since it is unlikely that randomly drawn locations will be placed in the key regions that reveal the difference. In this case, optimization of the test locations can considerably boost the test power (see ME-full in Figure 3.2d). Note also that SCF variants perform significantly better than ME variants on the Blobs problem, as the difference in P and Q is localized in the frequency domain; ME-full and ME-grid would require many more test locations in the spatial domain to match the test powers of the SCF variants. For the same reason, SCF-full does much better than the quadratic-time MMD across most sample sizes, as the latter represents a weighted distance between characteristic functions integrated across the entire frequency domain [Sriperumbudur et al., 2010, Corollary 4].

3.5.3 Test Power Vs. Dimension d

We next investigate how the dimension (d) of the problem can affect type-I errors and test powers of ME and SCF tests. We consider the same artificial problems: SG, GMD and GVD. This time, we fix the test sample size to 10000, set $J = 5$, and vary the dimension. The results are shown in Figure 3.3. Due to the large dimensions and sample size, it is computationally infeasible to run MMD-quad.

We observe that all the tests except the T-test can maintain type-I error at roughly the specified significance level $\alpha = 0.01$ as dimension increases. The type-I performance of the T-test is incorrect at large d because of the difficulty in accurately estimating the covariance matrix in high dimensions. As the T-test relies on only the

Table 3.2: Type-I errors and powers of various tests in the problem of distinguishing NIPS papers from two categories. $\alpha = 0.01$. $J = 1$. n_{te} denotes the test sample size of each of the two samples.

Problem	n^{te}	ME-full	ME-grid	SCF-full	SCF-grid	MMD-quad	MMD-lin
Bayes-Bayes	215	.012	.018	.012	.004	.022	.008
Bayes-Deep	216	.954	.034	.688	.180	.906	.262
Bayes-Learn	138	.990	.774	.836	.534	1.00	.238
Bayes-Neuro	394	1.00	.300	.828	.500	.952	.972
Learn-Deep	149	.956	.052	.656	.138	.876	.500
Learn-Neuro	146	.960	.572	.590	.360	1.00	.538

difference in the means, a difference of P and Q in the variance cannot be detected in the GVD problem (Figure 3.3c), even though the Gaussian assumptions hold. It is interesting to note the high performance of ME-full in the GMD problem in Figure 3.3b. ME-full achieves the maximum test power of 1.0 throughout and matches the power T-test, in spite of being nonparametric and making no assumption on P and Q (the T-test is further advantaged by its excessive Type-I error). However, this is true only with optimization of the test locations. This is reflected in the test power of ME-grid in Figure 3.3b which drops monotonically as dimension increases, highlighting the importance of test location optimization. The performance of MMD-lin degrades quickly with increasing dimension, as expected from Ramdas et al. [2015].

3.5.4 Distinguishing Articles From Two Categories

We now turn to performance on real data. We first consider the problem of distinguishing two categories of publications at the conference on Neural Information Processing Systems (NIPS). Out of 5903 papers published in NIPS from 1988 to 2015, we manually select disjoint subsets related to Bayesian inference (Bayes), neuroscience (Neuro), deep learning (Deep), and statistical learning theory (Learn) (see Section 3.7). Each paper is represented as a bag of words using TF-IDF [Manning et al., 2008] as features. We perform stemming, remove all stop words, and retain only nouns. A further filtering of document-frequency (DF) of words that satisfies $5 \leq \text{DF} \leq 2000$ yields approximately 5000 words from which 2000 words (i.e., $d = 2000$ dimensions) are randomly selected. See Section 3.7 for more details on the preprocessing. For ME and SCF tests, we use only one test location i.e., set $J = 1$. We perform 1000 permutations to approximate the null distribution of MMD-quad in this and the following experiments. We run the test 500 times where in each trial half of the collection is randomly assigned as a training set (D^{tr}) for tuning the test locations and the Gaussian width. The other half (D^{te}) is used for the actual two-sample test. If the number of papers in one sample is larger, we subsample so that the size matches the smaller sample.

Type-I errors and test powers are summarized in Table. 3.2. The first column indicates the categories of the papers in the two samples. In Bayes-Bayes problem, papers on Bayesian inference are randomly partitioned into two samples in each trial. This task represents a case in which H_0 holds. Among all the linear-time tests, we

observe that ME-full has the highest test power in all the tasks, attaining a maximum test power of 1.0 in the Bayes-Neuro problem. This high performance assures that although different test locations $\mathcal{V}$ may be selected in different trials, these locations are each informative. It is interesting to observe that ME-full has performance close to or better than MMD-quad, which requires $O(n^2)$ runtime complexity. Besides clear advantages of interpretability and linear runtime of the proposed tests, these results suggest that evaluating the differences in expectations of analytic functions at particular locations can yield an equally powerful test at a much lower cost, as opposed to computing the RKHS norm of the witness function as done in MMD. Unlike Blobs (Figure 3.2d), however, Fourier features are less powerful in this setting.

We further investigate the interpretability of the ME test by the following procedure. For the learned test location $\mathbf{v}^t \in \mathbb{R}^d$ ($d = 2000$) in trial t, we construct $\tilde{\mathbf{v}}^t = \left(\tilde{v}_1^t, \ldots, \tilde{v}_d^t\right)$ such that $\tilde{v}_j^t = |v_j^t|$. Let $\eta_j^t \in \{0, 1\}$ be an indicator variable taking value 1 if $\tilde{v}_j^t$ is among the top five largest for all $j \in \{1, \ldots, d\}$, and 0 otherwise. Define $\eta_j := \sum_t \eta_j^t$ as a proxy indicating the significance of word j i.e., η_j is high if word j is frequently among the top five largest as measured by $\tilde{v}_j^t$. The top ten words as sorted in descending order by η_j in all the problems are

- **Bayes-Bayes**: collabor, traffic, bay, permut, net, central, occlus, mask, draw, joint.
- **Bayes-Deep**: infer, bay, mont, adaptor, motif, haplotyp, ecg, covari, boltzmann, classifi.
- **Bayes-Learn**: infer, markov, graphic, segment, bandit, boundari, favor, carlo, prioriti, prop.
- **Bayes-Neuro**: spike, markov, cortex, dropout, recurr, iii, gibb, basin, circuit, subsystem.
- **Learn-Deep**: deep, forward, delay, subgroup, bandit, recept, invari, overlap, inequ, pia.
- **Learn-Neuro**: polici, interconnect, hardwar, decay, histolog, edg, period, basin, inject, human.

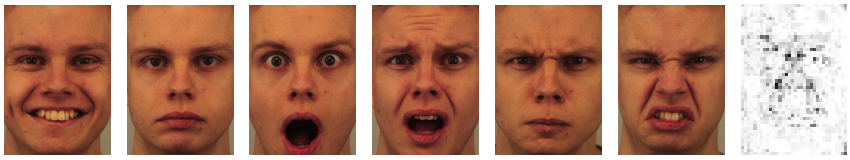
The results show that the learned test locations are highly interpretable. Indeed, in the Bayes-Neuro problem, "markov" and "gibb" (i.e., stemmed from Gibbs) are discriminative terms in Bayesian inference category, and "spike" and "cortex" are key terms in neuroscience. Note that since H_0 is true in the Bayes-Bayes problem, the extracted terms are arbitrary and not interpretable. To show that not all the randomly selected 2000 terms are informative, if the definition of η_j^t is modified to consider the least important words (i.e., η_j is high if word j is frequently among the top five smallest as measured by $\tilde{v}_j^t$), we instead obtain

circumfer, bra, dominiqu, rhino, mitra, kid, impostor,

in the Bayes-Neuro problem. These words are not discriminative.

3.5.5 Distinguishing Positive and Negative Emotions

In the final experiment, we study how well ME and SCF tests can distinguish two samples of photos of people showing positive and negative facial expressions. Our



(a) HA (b) NE (c) SU (d) AF (e) AN (f) DI (g) $\mathbf{v}_1$

Figure 3.4: (a)-(f): Six facial expressions of actor AM05 in the KDEF data. (g): Average across trials of the learned test locations $\mathbf{v}_1$.

emphasis is on the discriminative features of the faces identified by ME test showing how the two groups differ. For this purpose, we use Karolinska Directed Emotional Faces (KDEF) dataset [Lundqvist et al., 1998] containing aligned face images of 70 amateur actors, 35 females and 35 males. We use only photos showing front views of the faces. In the dataset, each actor displays seven expressions: happy (HA), neutral (NE), surprised (SU), sad (SA), afraid (AF), angry (AN), and disgusted (DI). We assign HA, NE, and SU faces into the positive emotion group (i.e., samples from P), and AF, AN and DI faces into the negative emotion group (samples from Q). We denote this problem as "$+$ vs. $-$". Examples of six facial expressions from one actor are shown in Figure 3.4. Photos of the SA group are unused to keep the sizes of the two samples the same. Each image of size 562×762 pixels is cropped to exclude the background, resized to $48 \times 34 = 1632$ pixels (d), and converted to grayscale.

We run the tests 500 times with the same setting used previously i.e., Gaussian kernels, and $J = 1$. The type-I errors and test powers are shown in Table 3.3. In the table, "$\pm$ vs. $\pm$" is a problem in which all faces expressing the six emotions are randomly split into two samples of equal sizes i.e., H_0 is true. Both ME-full and SCF-full achieve high test powers while maintaining the correct type-I errors.

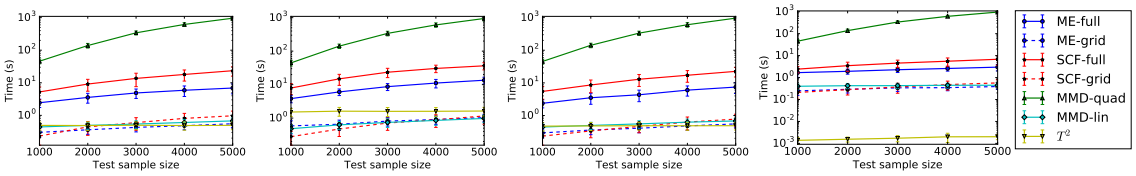
Table 3.3: Type-I errors and powers in the problem of distinguishing positive (+) and negative (-) facial expressions. $\alpha = 0.01$. $J = 1$.

Problem	n^{te}	ME-full	ME-grid	SCF-full	SCF-grid	MMD-quad	MMD-lin
$\pm$ vs. $\pm$	201	.010	.012	.014	.002	.018	.008
$+$ vs. $-$	201	.998	.656	1.00	.750	1.00	.578

As a way to interpret how positive and negative emotions differ, we take an average across trials of the learned test locations of ME-full in the "$+$ vs. $-$" problem. This average is shown in Figure 3.4g. We see that the test locations faithfully capture the difference of positive and negative emotions by giving more weights to the regions of nose, upper lip, and nasolabial folds (smile lines), confirming the interpretability of the test in a high-dimensional setting.

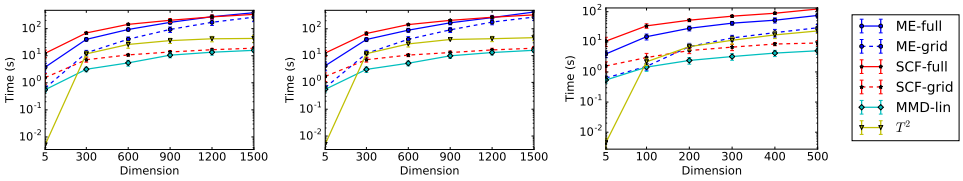
3.6 Runtimes

In this section, we provide runtimes of all the experiments. The runtimes of the "Test power vs. sample n" experiment are shown in Figure 3.5. The runtimes of the "Test power vs. dimension d" experiment are shown in Figure 3.6. Tables 3.4, 3.5 give the runtimes of the two real-data experiments.



(a) SG. $d = 50$. (b) GMD. $d = 100$. (c) GVD. $d = 50$. (d) Blobs.

Figure 3.5: Plots of runtimes in the "Test power vs. sample n" experiment.



(a) SG (b) GMD (c) GVD

Figure 3.6: Plots of runtimes in the "Test power vs. dimension d" experiment. The test sample size is 10000.

Table 3.4: Runtimes (in seconds) in the problem of distinguishing NIPS papers from two categories.

Problem	n^{te}	ME-full	ME-grid	SCF-full	SCF-grid	MMD-quad	MMD-lin
Bayes-Bayes	215	126.7	116	34.67	1.855	13.66	.6112
Bayes-Deep	216	118.3	111.7	36.41	1.933	13.59	.5105
Bayes-Learn	138	94.59	89.16	23.69	1.036	2.152	.36
Bayes-Neuro	394	142.5	130.3	69.19	3.533	32.71	.8643
Learn-Deep	149	105	99.59	24.99	1.253	2.417	.4744
Learn-Neuro	146	101.2	93.53	25.29	1.178	2.351	.3658

Table 3.5: Runtimes (in seconds) in the problem of distinguishing positive (+) and negative (-) facial expressions.

Problem	n^{te}	ME-full	ME-grid	SCF-full	SCF-grid	MMD-quad	MMD-lin
$\pm$ vs. $\pm$	201	87.7	83.4	10.5	1.45	9.93	0.464
$+$ vs. $-$	201	85.0	80.6	11.7	1.42	10.4	0.482

In the cases where n is large (Figure 3.5), MMD-quad has the largest runtime due to its quadratic dependency on the sample size. In the extreme case where the test

sample size is 10000 (Figure 3.6), it is computationally infeasible to run MMD-quad. We observe that the proposed ME-full and SCF-full have a slight overhead from the parameter optimization. However, since the optimization procedure is also linear in n, we are able to conduct an accurate test in less than 10 minutes even when the test sample size is 10000 and $d = 1500$ (see Figures 3.6a, 3.6b). We note that the actual tests (after optimization) for all ME and SCF variants take less than one second in all cases. In the ME-full, we initialize the test locations with realizations from two multivariate normal distributions fitted to samples from P and Q. When d is large, this heuristic can be expensive. An alternative initialization scheme for $\mathcal{V}$ is to randomly select J points from the two samples.

3.7 Preprocessing of the NIPS Text Collection

The full procedure for processing the NIPS text collection (in Section 3.5.4) is summarized as following.

1. Download all 5903 papers from 1988 to 2015 from `https://papers.nips.cc` as PDF files.
2. Convert each PDF file to text with `pdftotext`.[3]
3. Remove all stop words. We use the list of stop words from `http://www.ranks.nl/stopwords`.
4. Keep only nouns. We use the list of nouns as available in WordNet-3.0.[4]
5. Keep only words which contain only English alphabets i.e., does not contain punctuations or numbers. Also, word length must be between 3 and 20 characters (inclusive).
6. Keep only words which occur in at least 5 documents, and in no more than 2000 documents.
7. Convert all characters to small case. Stem all words with SnowballStemmer in NLTK [Bird et al., 2009]. For example, "recognize" and "recognizer" become "recogn" after stemming.
8. Categorize papers into disjoint collections. A paper is treated as belonging to a group if its title has at least one word from the list of keywords for the category. Papers that match the criteria of both categories are not considered. The lists of keywords are as follows.

    (a) **Bayesian inference** (Bayes): graphical model, bayesian, inference, mcmc, monte carlo, posterior, prior, variational, markov, latent, probabilistic, exponential family.

[3] `pdftotext` is available at `http://poppler.freedesktop.org`.
[4] WordNet is available online at `https://wordnet.princeton.edu/wordnet/citing-wordnet/`.

(b) **Deep learning** (Deep): deep, drop out, auto-encod, convolutional, neural net, belief net, boltzmann.

(c) **Learning theory** (Learn): learning theory, consistency, theoretical guarantee, complexity, pac-bayes, pac-learning, generalization, uniform converg, bound, deviation, inequality, risk min, minimax, structural risk, VC, rademacher, asymptotic.

(d) **Neuroscience** (Neuro): motor control, neural, neuron, spiking, spike, cortex, plasticity, neural decod, neural encod, brain imag, biolog, perception, cognitive, emotion, synap, neural population, cortical, firing rate, firing-rate, sensor.

9. Randomly select 2000 words from the remaining words.

10. Treat each paper as a bag of words and construct a feature vector with TF-IDF [Manning et al., 2008].

Proofs

3.A Proof: Convergence of the NME Power Criterion

Notations Let $\langle \mathbf{A}, \mathbf{B} \rangle_F := \operatorname{tr}\left(\mathbf{A}^\top \mathbf{B}\right)$ be the Frobenius inner product, and $\|\mathbf{A}\|_F := \sqrt{\langle \mathbf{A}, \mathbf{A} \rangle_F}$. $\mathbf{A} \succeq \mathbf{0}$ means that $\mathbf{A} \in \mathbb{R}^{d \times d}$ is symmetric, positive semidefinite. For $\mathbf{a} \in \mathbb{R}^d$, $\|\mathbf{a}\|_2^2 = \langle \mathbf{a}, \mathbf{a} \rangle_2 = \mathbf{a}^\top \mathbf{a}$. $[\mathbf{a}_1; \ldots; \mathbf{a}_N] \in \mathbb{R}^{d_1 + \ldots + d_N}$ is the concatenation of the $\mathbf{a}_n \in \mathbb{R}^{d_n}$ vectors. $\mathbb{R}^+$ is the set of positive reals. $f \circ g$ is the composition of function f and g. Let $\mathcal{M}$ denote a general metric space below. In measurability requirements metric spaces are meant to be endowed with their Borel σ-algebras.

Let $\mathcal{C}$ be a collection of subsets of $\mathcal{M}$ ($C \subseteq 2^{\mathcal{M}}$). $\mathcal{C}$ is said to shatter $\{p_1, p_2, \ldots, p_i\} \subseteq \mathcal{M}$ set, if for any $S \subseteq \{p_1, p_2, \ldots, p_i\}$ there exist $C \in \mathcal{C}$ such that $S = C \cap \{p_1, p_2, \ldots, p_i\}$; in other words, arbitrary subset of $\{p_1, p_2, \ldots, p_i\}$ can be cut out by an element of $\mathcal{C}$. The VC index of $\mathcal{C}$ is the smallest i for which no set of size i is shattered:

$$VC\left(\mathcal{C}\right) = \inf\left\{ i : \max_{p_1, \ldots, p_i} \left|\left\{ C \cap \{p_1, \ldots, p_i\} : C \in \mathcal{C} \right\}\right| < 2^i \right\}.$$

A collection $\mathcal{C}$ of measurable sets is called VC-class if its index $VC\left(\mathcal{C}\right)$ is finite. The subgraph of a real-valued function $f : \mathcal{M} \to \mathbb{R}$ is $sub(f) = \{(m, u) : u < f(m)\} \subseteq \mathcal{M} \times \mathbb{R}$. A collection of $\mathcal{F}$ measurable functions is called VC-subgraph class, or shortly VC if the collection of all subgraphs of $\mathcal{F}$, $\{sub(f)\}_{f \in \mathcal{F}}$ is a VC-class of sets; its index is defined as $VC\left(\mathcal{F}\right) := VC\left(\{sub(f)\}_{f \in \mathcal{F}}\right)$.

Let $L^0(\mathcal{M})$ be the set of $\mathcal{M} \to \mathbb{R}$ measurable functions. Given an i.i.d. (independent identically distributed) sample from $\mathbb{P}$ ($w_i \overset{i.i.d.}{\sim} \mathbb{P}$), let $w_{1:n} = (w_1, \ldots, w_n)$ and let $\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{w_i}$ denote the empirical measure.

$$L^q\left(\mathcal{M}, \mathbb{P}_n\right) = \left\{ f \in L^0\left(\mathcal{M}\right) : \|f\|_{L^q(\mathcal{M}, \mathbb{P}_n)} = \left[\frac{1}{n} \sum_{i=1}^n |f(w_i)|^q \right]^{\frac{1}{q}} < \infty \right\},$$

where $(1 \leq q < \infty)$, and $\|f\|_{L^\infty(\mathcal{M})} := \sup_{m \in \mathcal{M}} |f(m)|$.

Define $\mathbb{P}f := \int_{\mathcal{M}} f(w) \mathrm{d}\mathbb{P}(w)$, where $\mathbb{P}$ is a probability distribution on $\mathcal{M}$. Let $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - \mathbb{P}f|$. The diameter of a class $\mathcal{F} \subseteq L^2\left(\mathcal{M}, \mathbb{P}_n\right)$ is

$$\operatorname{diam}(\mathcal{F}, L^2(\mathcal{M}, \mathbb{P}_n)) := \sup_{f, f' \in \mathcal{F}} \|f - f'\|_{L^2(\mathcal{M}, \mathbb{P}_n)},$$

its r-covering number ($r > 0$) is the size of the smallest r-net

$$N\left(r, \mathcal{F}, L^2\left(\mathcal{M}, \mathbb{P}_n\right)\right) = \inf\left\{t \geq 1 : \exists f_1, \ldots, f_t \in \mathcal{F} \text{ such that } \mathcal{F} \subseteq \cup_{i=1}^{t} B(r, f_i)\right\},$$

where $B(r,f) = \left\{g \in L^2\left(\mathcal{M}, \mathbb{P}_n\right) \mid \|f-g\|_{L^2(\mathcal{M},\mathbb{P}_n)} \leq r\right\}$ is the ball with center f and radius r. $\times_{i=1}^{N}\mathbb{Q}_i$ is the N-fold product measure. For sets Q_i, $\times_{i=1}^{n}Q_i$ is their Cartesian product. For a function class $\mathcal{F} \subseteq L^0(\mathcal{M})$ and $w_{1:n} \in \mathcal{M}^n$, $R(\mathcal{F}, w_{1:n}) := \mathbb{E}_{\mathbf{r}}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} r_i f(w_i)\right|\right]$ is the empirical Rademacher average, where $\mathbf{r} := r_{1:n}$ and r_i-s are i.i.d. samples from a Rademacher random variable [$\mathbb{P}(r_i = 1) = \mathbb{P}(r_i = -1) = \frac{1}{2}$]. Let (Θ, ρ) be a metric space; a collection of $\mathcal{F} = \{f_\theta \mid \theta \in \Theta\} \subseteq L^0(\mathcal{M})$ functions is called a separable Carathéodory family if Θ is separable and $\theta \mapsto f_\theta(m)$ is continuous for all $m \in \mathcal{M}$. $span(\cdot)$ denotes the linear hull of its arguments. $\Gamma(t) = \int_0^\infty u^{t-1}e^{-u}\mathrm{d}u$ denotes the Gamma function.

Recall Theorem 3.6:

Theorem 3.6 (Convergence of the power criterion of the NME test)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a measurable set, and $\mathbb{V}$ be a collection in which each element is a set of J test locations. For a class of kernels $\mathcal{K}$ on $\mathcal{X} \subseteq \mathbb{R}^d$, define*

$$\mathcal{F}_1 := \{\mathbf{x} \mapsto k(\mathbf{x}, \mathbf{v}) \mid k \in \mathcal{K}, \mathbf{v} \in \mathcal{X}\}, \tag{3.11}$$
$$\mathcal{F}_2 := \{\mathbf{x} \mapsto k(\mathbf{x}, \mathbf{v})k(\mathbf{x}, \mathbf{v}') \mid k \in \mathcal{K}, \mathbf{v}, \mathbf{v}' \in \mathcal{X}\},$$
$$\mathcal{F}_3 := \{(\mathbf{x}, \mathbf{y}) \mapsto k(\mathbf{x}, \mathbf{v})k(\mathbf{y}, \mathbf{v}') \mid k \in \mathcal{K}, \mathbf{v}, \mathbf{v}' \in \mathcal{X}\}. \tag{3.12}$$

Assume

1. *$\mathcal{K}$ is a uniformly bounded (by B) family of $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ measurable kernels,*
2. *$\tilde{c} := \sup_{\mathcal{V}\in\mathbb{V}} \sup_{k\in\mathcal{K}} \|\mathbf{\Sigma}^{-1}\|_F < \infty$, and*
3. *$\mathcal{F}_i = \{f_{\theta_i} \mid \theta_i \in \Theta_i\}$ is VC-subgraph [van der Vaart and Wellner, 2000] with VC-index $VC(\mathcal{F}_i) < \infty$, and $\mathcal{F}_i$ is a separable Carathéodory family (i.e., Θ_i is in a separable metric space, and $\theta_i \mapsto f_{\theta_i}(\mathbf{x})$ is continuous for all $\mathbf{x} \in \mathcal{X}$), for all $i = 1, 2, 3$.*

Let $\bar{c}_1 := 4B^2J\sqrt{J}\tilde{c}, \bar{c}_2 := 4B\sqrt{J}\tilde{c}$, and $\bar{c}_3 := 4B^2J\tilde{c}^2$. Let C_i-s ($i = 1, 2, 3$) be the universal constants associated to $\mathcal{F}_i$-s according to Theorem 2.6.7 in van der Vaart and Wellner [2000]. Then for any $\delta \in (0, 1)$ with probability at least $1 - \delta$,

$$\sup_{\mathcal{V}\in\mathbb{V}} \sup_{k\in\mathcal{K}} \left|\bar{\mathbf{z}}_n^\top (\mathbf{S}_n + \gamma_n \mathbf{I})^{-1}\bar{\mathbf{z}}_n - \boldsymbol{\mu}^\top \mathbf{\Sigma}^{-1}\boldsymbol{\mu}\right|$$
$$\leq 2T_{\mathcal{F}_1}\left(\frac{2}{\gamma_n}\bar{c}_1 BJ\frac{2n-1}{n-1} + \bar{c}_2\sqrt{J}\right) + \frac{2}{\gamma_n}\bar{c}_1 J(T_{\mathcal{F}_2} + T_{\mathcal{F}_3}) + \frac{8}{\gamma_n}\frac{\bar{c}_1 B^2 J}{n-1} + \bar{c}_3\gamma_n,$$

where

$$T_{\mathcal{F}_j} = \frac{16\sqrt{2}B^{\zeta_j}}{\sqrt{n}}\left(2\sqrt{\log\left[C_j \times VC(\mathcal{F}_j)(16e)^{VC(\mathcal{F}_j)}\right]} + \frac{\sqrt{2\pi[VC(\mathcal{F}_j) - 1]}}{2}\right) + B^{\zeta_j}\sqrt{\frac{2\log(5/\delta)}{n}},$$

for $j = 1, 2, 3$ *and* $\zeta_1 = 1, \zeta_2 = \zeta_3 = 2$.

A proof is given as follows.

3.A.1 Bound in Terms of $\mathbf{S}_n$ and $\overline{\mathbf{z}}_n$

For brevity, we will interchangeably use $\mathbf{S}_n$ for $\mathbf{S}_n(\mathcal{V})$ and $\overline{\mathbf{z}}_n$ for $\overline{\mathbf{z}}_n(\mathcal{V})$, and write $\sup_{\mathcal{V},k}$ for $\sup_{\mathcal{V}\in\mathbb{V}} \sup_{k\in\mathcal{K}}$. $\mathbf{S}_n(\mathcal{V})$ and $\overline{\mathbf{z}}_n(\mathcal{V})$ will be used mainly when the dependency of $\mathcal{V}$ needs to be emphasized. All suprema over $\mathcal{V}$ and the kernel k should be understood as being constrained within $\mathbb{V}$ and $\mathcal{K}$ respectively. We start with

$$\sup_{\mathcal{V},k} \left| \overline{\mathbf{z}}_n^\top (\mathbf{S}_n + \gamma_n I)^{-1} \overline{\mathbf{z}}_n - \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right|,$$

and upper bound the argument of $\sup_{\mathcal{V},k}$ as

$$\begin{aligned}
&\left| \overline{\mathbf{z}}_n^\top (\mathbf{S}_n + \gamma_n I)^{-1} \overline{\mathbf{z}}_n - \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right| \\
&= \left| \overline{\mathbf{z}}_n^\top (\mathbf{S}_n + \gamma_n I)^{-1} \overline{\mathbf{z}}_n - \boldsymbol{\mu}^\top (\boldsymbol{\Sigma} + \gamma_n I)^{-1} \boldsymbol{\mu} + \boldsymbol{\mu}^\top (\boldsymbol{\Sigma} + \gamma_n I)^{-1} \boldsymbol{\mu} - \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right| \\
&\leq \left| \overline{\mathbf{z}}_n^\top (\mathbf{S}_n + \gamma_n I)^{-1} \overline{\mathbf{z}}_n - \boldsymbol{\mu}^\top (\boldsymbol{\Sigma} + \gamma_n I)^{-1} \boldsymbol{\mu} \right| + \left| \boldsymbol{\mu}^\top (\boldsymbol{\Sigma} + \gamma_n I)^{-1} \boldsymbol{\mu} - \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right| \\
&:= (\square_1) + (\square_2).
\end{aligned}$$

For $(\square_1)$, we have

$$\begin{aligned}
&\left| \overline{\mathbf{z}}_n^\top (\mathbf{S}_n + \gamma_n I)^{-1} \overline{\mathbf{z}}_n - \boldsymbol{\mu}^\top (\boldsymbol{\Sigma} + \gamma_n I)^{-1} \boldsymbol{\mu} \right| \\
&= \left| \left\langle \overline{\mathbf{z}}_n \overline{\mathbf{z}}_n^\top, (\mathbf{S}_n + \gamma_n I)^{-1} \right\rangle_F - \left\langle \boldsymbol{\mu}\boldsymbol{\mu}^\top, (\boldsymbol{\Sigma} + \gamma_n I)^{-1} \right\rangle_F \right| \\
&= \left| \left\langle \overline{\mathbf{z}}_n \overline{\mathbf{z}}_n^\top, (\mathbf{S}_n + \gamma_n I)^{-1} \right\rangle_F - \left\langle \overline{\mathbf{z}}_n \overline{\mathbf{z}}_n^\top, (\boldsymbol{\Sigma} + \gamma_n I)^{-1} \right\rangle_F + \left\langle \overline{\mathbf{z}}_n \overline{\mathbf{z}}_n^\top, (\boldsymbol{\Sigma} + \gamma_n I)^{-1} \right\rangle_F - \left\langle \boldsymbol{\mu}\boldsymbol{\mu}^\top, (\boldsymbol{\Sigma} + \gamma_n I)^{-1} \right\rangle_F \right| \\
&\leq \left| \left\langle \overline{\mathbf{z}}_n \overline{\mathbf{z}}_n^\top, (\mathbf{S}_n + \gamma_n I)^{-1} - (\boldsymbol{\Sigma} + \gamma_n I)^{-1} \right\rangle_F \right| + \left| \left\langle \overline{\mathbf{z}}_n \overline{\mathbf{z}}_n^\top - \boldsymbol{\mu}\boldsymbol{\mu}^\top, (\boldsymbol{\Sigma} + \gamma_n I)^{-1} \right\rangle_F \right| \\
&= \|\overline{\mathbf{z}}_n \overline{\mathbf{z}}_n^\top\|_F \| (\mathbf{S}_n + \gamma_n I)^{-1} - (\boldsymbol{\Sigma} + \gamma_n I)^{-1} \|_F + \|\overline{\mathbf{z}}_n \overline{\mathbf{z}}_n^\top - \boldsymbol{\mu}\boldsymbol{\mu}^\top\|_F \| (\boldsymbol{\Sigma} + \gamma_n I)^{-1} \|_F \\
&\overset{(a)}{\leq} \|\overline{\mathbf{z}}_n \overline{\mathbf{z}}_n^\top\|_F \| (\mathbf{S}_n + \gamma_n I)^{-1} [(\boldsymbol{\Sigma} + \gamma_n I) - (\mathbf{S}_n + \gamma_n I)] (\boldsymbol{\Sigma} + \gamma_n I)^{-1} \|_F \\
&\quad + \|\overline{\mathbf{z}}_n \overline{\mathbf{z}}_n^\top - \overline{\mathbf{z}}_n \boldsymbol{\mu}^\top + \overline{\mathbf{z}}_n \boldsymbol{\mu}^\top - \boldsymbol{\mu}\boldsymbol{\mu}^\top\|_F \|\boldsymbol{\Sigma}^{-1}\|_F \\
&\overset{(a)}{\leq} \|\overline{\mathbf{z}}_n \overline{\mathbf{z}}_n^\top\|_F \|(\mathbf{S}_n + \gamma_n I)^{-1}\|_F \|\boldsymbol{\Sigma} - \mathbf{S}_n\|_F \|\boldsymbol{\Sigma}^{-1}\|_F + \|\overline{\mathbf{z}}_n (\overline{\mathbf{z}}_n - \boldsymbol{\mu})^\top\|_F \|\boldsymbol{\Sigma}^{-1}\|_F + \|(\overline{\mathbf{z}}_n - \boldsymbol{\mu})\boldsymbol{\mu}^\top\|_F \|\boldsymbol{\Sigma}^{-1}\|_F \\
&\overset{(b)}{\leq} \frac{\sqrt{J}}{\gamma_n} \|\overline{\mathbf{z}}_n\|_2^2 \|\boldsymbol{\Sigma} - \mathbf{S}_n\|_F \|\boldsymbol{\Sigma}^{-1}\|_F + \|\overline{\mathbf{z}}_n\|_2 \|\overline{\mathbf{z}}_n - \boldsymbol{\mu}\|_2 \|\boldsymbol{\Sigma}^{-1}\|_F + \|\boldsymbol{\mu}\|_2 \|\overline{\mathbf{z}}_n - \boldsymbol{\mu}\|_2 \|\boldsymbol{\Sigma}^{-1}\|_F,
\end{aligned}$$

where at (a) we use $\| (\boldsymbol{\Sigma} + \gamma_n I)^{-1} \|_F \leq \|\boldsymbol{\Sigma}^{-1}\|_F$ and at (b) we use $\|(\mathbf{S}_n + \gamma_n I)^{-1}\|_F \leq \sqrt{J}\|(\mathbf{S}_n + \gamma_n I)^{-1}\|_2 \leq \sqrt{J}/\gamma_n$.

For $(\square_2)$, we have

$$\begin{aligned}
\left| \boldsymbol{\mu}^\top (\boldsymbol{\Sigma} + \gamma_n I)^{-1} \boldsymbol{\mu} - \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right| &= \left| \left\langle \boldsymbol{\mu}\boldsymbol{\mu}^\top, (\boldsymbol{\Sigma} + \gamma_n I)^{-1} - \boldsymbol{\Sigma}^{-1} \right\rangle_F \right| \\
&\leq \|\boldsymbol{\mu}\boldsymbol{\mu}^\top\|_F \| (\boldsymbol{\Sigma} + \gamma_n I)^{-1} - \boldsymbol{\Sigma}^{-1} \|_F \\
&= \|\boldsymbol{\mu}\|_2^2 \| (\boldsymbol{\Sigma} + \gamma_n I)^{-1} [\boldsymbol{\Sigma} - (\boldsymbol{\Sigma} + \gamma_n I)] \boldsymbol{\Sigma}^{-1} \|_F
\end{aligned}$$

$$
\begin{aligned}
&= \gamma_n \|\boldsymbol{\mu}\|_2^2 \| \left(\boldsymbol{\Sigma} + \gamma_n I\right)^{-1} \boldsymbol{\Sigma}^{-1} \|_F \\
&\leq \gamma_n \|\boldsymbol{\mu}\|_2^2 \| \left(\boldsymbol{\Sigma} + \gamma_n I\right)^{-1} \|_F \|\boldsymbol{\Sigma}^{-1}\|_F \\
&\overset{(a)}{\leq} \gamma_n \|\boldsymbol{\mu}\|_2^2 \|\boldsymbol{\Sigma}^{-1}\|_F^2.
\end{aligned}
$$

Combining the upper bounds for $(\square_1)$ and $(\square_2)$, we arrive at

$$
\begin{aligned}
&\left| \bar{\mathbf{z}}_n^\top (\mathbf{S}_n + \gamma_n I)^{-1} \bar{\mathbf{z}}_n - \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right| \\
&\leq \frac{\sqrt{J}}{\gamma_n} \|\bar{\mathbf{z}}_n\|_2^2 \|\boldsymbol{\Sigma} - \mathbf{S}_n\|_F \|\boldsymbol{\Sigma}^{-1}\|_F + (\|\bar{\mathbf{z}}_n\|_2 + \|\boldsymbol{\mu}\|_2) \|\bar{\mathbf{z}}_n - \boldsymbol{\mu}\|_2 \|\boldsymbol{\Sigma}^{-1}\|_F + \gamma_n \|\boldsymbol{\mu}\|_2^2 \|\boldsymbol{\Sigma}^{-1}\|_F^2 \\
&\leq 4B^2 J \tilde{c} \frac{\sqrt{J}}{\gamma_n} \|\boldsymbol{\Sigma} - \mathbf{S}_n\|_F + 4B\sqrt{J}\tilde{c} \|\bar{\mathbf{z}}_n - \boldsymbol{\mu}\|_2 + 4B^2 J \tilde{c}^2 \gamma_n \\
&= \frac{\bar{c}_1}{\gamma_n} \|\boldsymbol{\Sigma} - \mathbf{S}_n\|_F + \bar{c}_2 \|\bar{\mathbf{z}}_n - \boldsymbol{\mu}\|_2 + \bar{c}_3 \gamma_n
\end{aligned} \tag{3.14}
$$

with $\bar{c}_1 := 4B^2 J\sqrt{J}\tilde{c}, \bar{c}_2 := 4B\sqrt{J}\tilde{c}$, $\bar{c}_3 := 4B^2 J\tilde{c}^2$, and $\tilde{c} := \sup_{\mathcal{V},k} \|\boldsymbol{\Sigma}^{-1}\|_F < \infty$, where we applied the triangle inequality, the CBS (Cauchy-Bunyakovskii-Schwarz) inequality, and $\|\mathbf{a}\mathbf{b}^\top\|_F = \|\mathbf{a}\|_2 \|\mathbf{b}\|_2$. The boundedness of kernel k with the Jensen inequality implies that

$$
\begin{aligned}
\|\bar{\mathbf{z}}_n\|_2^2 = \|\frac{1}{n}\sum_{i=1}^n \mathbf{z}_i\|_2^2 \leq \frac{1}{n}\sum_{i=1}^n \|\mathbf{z}_i\|_2^2 &= \frac{1}{n}\sum_{i=1}^n \|\left(k(\mathbf{x}_i, \mathbf{v}_j) - k(\mathbf{y}_i, \mathbf{v}_j)\right)_{j=1}^J\|_2^2 \qquad (3.15) \\
&= \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^J \left[k(\mathbf{x}_i, \mathbf{v}_j) - k(\mathbf{y}_i, \mathbf{v}_j)\right]^2 \\
&\leq \frac{2}{n}\sum_{i=1}^n \sum_{j=1}^J k^2(\mathbf{x}_i, \mathbf{v}_j) + k^2(\mathbf{y}_i, \mathbf{v}_j) \leq 4B^2 J, \qquad (3.16)
\end{aligned}
$$

$$
\|\boldsymbol{\mu}(\mathcal{V})\|_2^2 = \sum_{j=1}^J \left(\mathbb{E}_{\mathbf{xy}}\left[k(\mathbf{x}, \mathbf{v}_j) - k(\mathbf{y}, \mathbf{v}_j)\right]\right)^2 \leq \sum_{j=1}^J \mathbb{E}_{\mathbf{xy}}\left[k(\mathbf{x}, \mathbf{v}_j) - k(\mathbf{y}, \mathbf{v}_j)\right]^2 \leq 4B^2 J. \tag{3.17}
$$

Taking sup in (3.14), we get

$$
\sup_{\mathcal{V},k} \left| \bar{\mathbf{z}}_n^\top (\mathbf{S}_n + \gamma_n I)^{-1} \bar{\mathbf{z}}_n - \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right| \leq \frac{\bar{c}_1}{\gamma_n} \sup_{\mathcal{V},k} \|\boldsymbol{\Sigma} - \mathbf{S}_n\|_F + \bar{c}_2 \sup_{\mathcal{V},k} \|\bar{\mathbf{z}}_n - \boldsymbol{\mu}\|_2 + \bar{c}_3 \gamma_n.
$$

3.A.2 Empirical Process Bound on $\bar{\mathbf{z}}_n$

Recall that $\bar{\mathbf{z}}_n(\mathcal{V}) = \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i(\mathcal{V}) \in \mathbb{R}^J$, $\mathbf{z}_i(\mathcal{V}) = \left(k(\mathbf{x}_i, \mathbf{v}_j) - k(\mathbf{y}_i, \mathbf{v}_j)\right)_{j=1}^J \in \mathbb{R}^J$, $\boldsymbol{\mu}(\mathcal{V}) = \left(\mathbb{E}_{\mathbf{xy}}\left[k(\mathbf{x}, \mathbf{v}_j) - k(\mathbf{y}, \mathbf{v}_j)\right]\right)_{j=1}^J$; thus

$$
\sup_{\mathcal{V}} \sup_{k \in \mathcal{K}} \|\bar{\mathbf{z}}_n(\mathcal{V}) - \boldsymbol{\mu}(\mathcal{V})\|_2 = \sup_{\mathcal{V}} \sup_{k \in \mathcal{K}} \sup_{\mathbf{b} \in B(\mathbf{1},\mathbf{0})} \langle \mathbf{b}, \bar{\mathbf{z}}_n(\mathcal{V}) - \boldsymbol{\mu}(\mathcal{V}) \rangle_2
$$

using that $\|\mathbf{a}\|_2 = \sup_{\mathbf{b}\in B(1,\mathbf{0})} \langle \mathbf{a}, \mathbf{b}\rangle_2$. Let us bound the argument of the supremum:

$$
\begin{aligned}
&\langle \mathbf{b}, \bar{\mathbf{z}}_n(\mathcal{V}) - \boldsymbol{\mu}(\mathcal{V})\rangle_2 \\
&\leq \sum_{j=1}^{J} |b_j| \left| \frac{1}{n}\sum_{i=1}^{n} \left[k(\mathbf{x}_i, \mathbf{v}_j) - k(\mathbf{y}_i, \mathbf{v}_j)\right] - \mathbb{E}_{\mathbf{xy}}\left[k(\mathbf{x}, \mathbf{v}_j) - k(\mathbf{y}, \mathbf{v}_j)\right]\right| \\
&\leq \sum_{j=1}^{J} |b_j| \left(\left| \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v}_j) - \mathbb{E}_{\mathbf{x}} k(\mathbf{x}, \mathbf{v}_j)\right| + \left| \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{y}_i, \mathbf{v}_j) - \mathbb{E}_{\mathbf{y}} k(\mathbf{y}, \mathbf{v}_j)\right| \right) \\
&\leq \sqrt{J} \sup_{\mathbf{v}\in\mathcal{X}} \sup_{k\in\mathcal{K}} \left| \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v}) - \mathbb{E}_{\mathbf{x}} k(\mathbf{x}, \mathbf{v})\right| + \sqrt{J} \sup_{\mathbf{v}\in\mathcal{X}} \sup_{k\in\mathcal{K}} \left| \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{y}_i, \mathbf{v}) - \mathbb{E}_{\mathbf{y}} k(\mathbf{y}, \mathbf{v})\right| \\
&= \sqrt{J}\, \|P_n - P\|_{\mathcal{F}_1} + \sqrt{J}\, \|Q_n - Q\|_{\mathcal{F}_1}
\end{aligned} \tag{3.18}
$$

by the triangle inequality and exploiting that $\|\mathbf{b}\|_1 \leq \sqrt{J}\,\|\mathbf{b}\|_2 \leq \sqrt{J}$ with $\mathbf{b} \in B(1, \mathbf{0})$. Thus, we have

$$
\sup_{\mathcal{V}} \sup_{k\in\mathcal{K}} \|\bar{\mathbf{z}}_n(\mathcal{V}) - \boldsymbol{\mu}(\mathcal{V})\|_2 \leq \sqrt{J}\, \|P_n - P\|_{\mathcal{F}_1} + \sqrt{J}\, \|Q_n - Q\|_{\mathcal{F}_1}.
$$

3.A.3 Empirical Process Bound on $\mathbf{S}_n$

Noting that

$$
\begin{aligned}
\boldsymbol{\Sigma}(\mathcal{V}) &= \mathbb{E}_{\mathbf{xy}}\left[\mathbf{z}(\mathcal{V})\mathbf{z}^\top(\mathcal{V})\right] - \boldsymbol{\mu}(\mathcal{V})\boldsymbol{\mu}^\top(\mathcal{V}), \\
\mathbf{S}_n(\mathcal{V}) &= \frac{1}{n}\sum_{a=1}^{n} \mathbf{z}_a(\mathcal{V})\mathbf{z}_a^\top(\mathcal{V}) - \frac{1}{n(n-1)}\sum_{a=1}^{n}\sum_{b\neq a} \mathbf{z}_a \mathbf{z}_b^T, \\
\mathbb{E}_{\mathbf{xy}}\left[\mathbf{z}(\mathcal{V})\mathbf{z}^\top(\mathcal{V})\right] &= \mathbb{E}_{\mathbf{xy}}\left[\frac{1}{n}\sum_{a=1}^{n} \mathbf{z}_a(\mathcal{V})\mathbf{z}_a^\top(\mathcal{V})\right], \\
\boldsymbol{\mu}(\mathcal{V})\boldsymbol{\mu}^\top(\mathcal{V}) &= \mathbb{E}_{\mathbf{xy}}\left[\frac{1}{n(n-1)}\sum_{a=1}^{n}\sum_{b\neq a} \mathbf{z}_a(\mathcal{V})\mathbf{z}_b^T(\mathcal{V})\right],
\end{aligned}
$$

we bound our target quantity as

$$
\begin{aligned}
&\|\mathbf{S}_n(\mathcal{V}) - \boldsymbol{\Sigma}(\mathcal{V})\|_F \\
&\leq \left\| \frac{1}{n}\sum_{a=1}^{n} \mathbf{z}_a(\mathcal{V})\mathbf{z}_a^\top(\mathcal{V}) - \mathbb{E}_{\mathbf{xy}}\left[\mathbf{z}(\mathcal{V})\mathbf{z}^\top(\mathcal{V})\right]\right\|_F + \left\| \frac{1}{n(n-1)}\sum_{a=1}^{n}\sum_{b\neq a} \mathbf{z}_a(\mathcal{V})\mathbf{z}_b^T(\mathcal{V}) - \boldsymbol{\mu}(\mathcal{V})\boldsymbol{\mu}^\top(\mathcal{V})\right\|_F \\
&=: (*_1) + (*_2).
\end{aligned} \tag{3.19}
$$

$$
\begin{aligned}
(*_2) &= \left\| \frac{1}{n}\sum_{a=1}^{n} \mathbf{z}_a(\mathcal{V}) \left[\frac{1}{n-1}\sum_{b\neq a} \mathbf{z}_b^\top(\mathcal{V})\right] - \boldsymbol{\mu}(\mathcal{V})\boldsymbol{\mu}^\top(\mathcal{V})\right\|_F \\
&\leq \left\| \frac{1}{n}\sum_{a=1}^{n} \mathbf{z}_a(\mathcal{V}) \left(\frac{1}{n-1}\sum_{b\neq a} \mathbf{z}_b^\top(\mathcal{V}) - \boldsymbol{\mu}^\top(\mathcal{V})\right)\right\|_F + \left\| \left(\frac{1}{n}\sum_{a=1}^{n} \mathbf{z}_a(\mathcal{V}) - \boldsymbol{\mu}(\mathcal{V})\right)\boldsymbol{\mu}^\top(\mathcal{V})\right\|_F
\end{aligned}
$$

$$
\begin{aligned}
&\leq \left\| \left(\frac{1}{n} \sum_{a=1}^{n} \mathbf{z}_a(\mathcal{V}) \right) \left(\frac{1}{n-1} \sum_{b=1}^{n} \mathbf{z}_b(\mathcal{V}) - \boldsymbol{\mu}(\mathcal{V}) \right)^\top \right\|_F + \left\| \left(\frac{1}{n} \sum_{a=1}^{n} \mathbf{z}_a(\mathcal{V}) \right) \frac{\mathbf{z}_a^\top(\mathcal{V})}{n-1} \right\|_F \\
&\quad + \left\| \left(\frac{1}{n} \sum_{a=1}^{n} \mathbf{z}_a(\mathcal{V}) - \boldsymbol{\mu}(\mathcal{V}) \right) \boldsymbol{\mu}^\top(\mathcal{V}) \right\|_F \\
&= \|\bar{\mathbf{z}}_n(\mathcal{V})\|_2 \left\| \frac{1}{n-1} \sum_{b=1}^{n} \mathbf{z}_b(\mathcal{V}) - \boldsymbol{\mu}(\mathcal{V}) \right\|_2 + \frac{1}{n-1} \|\bar{\mathbf{z}}_n(\mathcal{V})\|_2 \|\mathbf{z}_a(\mathcal{V})\|_2 \\
&\quad + \|\bar{\mathbf{z}}_n(\mathcal{V}) - \boldsymbol{\mu}(\mathcal{V})\|_2 \|\boldsymbol{\mu}(\mathcal{V})\|_2 \\
&\leq 2B\sqrt{J} \left(\frac{n}{n-1} \|\bar{\mathbf{z}}_n - \boldsymbol{\mu}(\mathcal{V})\|_2 + \frac{2B\sqrt{J}}{n-1} \right) + \frac{1}{n-1} 4B^2 J + 2B\sqrt{J} \|\bar{\mathbf{z}}_n(\mathcal{V}) - \boldsymbol{\mu}(\mathcal{V})\|_2 \\
&= \frac{8B^2 J}{n-1} + 2B\sqrt{J} \frac{2n-1}{n-1} \|\bar{\mathbf{z}}_n - \boldsymbol{\mu}(\mathcal{V})\|_2 \qquad (3.20)
\end{aligned}
$$

using the triangle inequality, the sub-additivity of sup, $\left\|\mathbf{a}\mathbf{b}^T\right\|_F = \|\mathbf{a}\|_2 \|\mathbf{b}\|_2$, $\|\bar{\mathbf{z}}_n(\mathcal{V})\|_2 \leq 2B\sqrt{J}$, $\|\mathbf{z}_a(\mathcal{V})\|_2 \leq 2B\sqrt{J}$ [see Eq. (3.16)] and

$$
\begin{aligned}
\left\| \frac{1}{n-1} \sum_{b=1}^{n} \mathbf{z}_b(\mathcal{V}) - \boldsymbol{\mu}(\mathcal{V}) \right\|_2 &= \left\| \frac{n}{n-1} \bar{\mathbf{z}}_n - \frac{n}{n-1} \boldsymbol{\mu}(\mathcal{V}) + \frac{1}{n-1} \boldsymbol{\mu}(\mathcal{V}) \right\|_2 \\
&\leq \frac{n}{n-1} \|\bar{\mathbf{z}}_n - \boldsymbol{\mu}(\mathcal{V})\|_2 + \frac{1}{n-1} \|\boldsymbol{\mu}(\mathcal{V})\|_2 ,
\end{aligned}
$$

with Eq. (3.17). Considering the first term in Eq. (3.19)

$$
\left\| \frac{1}{n} \sum_{a=1}^{n} \mathbf{z}_a \mathbf{z}_a^\top - \mathbb{E}_{\mathbf{xy}} \left[\mathbf{z}\mathbf{z}^\top \right] \right\|_F = \sup_{\mathbf{B} \in B(1,\mathbf{0})} \left\langle \mathbf{B}, \frac{1}{n} \sum_{a=1}^{n} \mathbf{z}_a \mathbf{z}_a^\top - \mathbb{E}_{\mathbf{xy}} \left[\mathbf{z}\mathbf{z}^\top \right] \right\rangle_F
$$

$$
\begin{aligned}
&\leq \sup_{\mathbf{B} \in B(1,\mathbf{0})} \sum_{i,j=1}^{J} |B_{ij}| \left| \frac{1}{n} \sum_{a=1}^{n} [k(\mathbf{x}_a, \mathbf{v}_i) - k(\mathbf{y}_a, \mathbf{v}_i)][k(\mathbf{x}_a, \mathbf{v}_j) - k(\mathbf{y}_a, \mathbf{v}_j)] \right. \\
&\quad \left. - \mathbb{E}_{\mathbf{xy}} \left([k(\mathbf{x}, \mathbf{v}_i) - k(\mathbf{y}, \mathbf{v}_i)] \left[k(\mathbf{x}, \mathbf{v}_j) - k(\mathbf{y}, \mathbf{v}_j) \right] \right) \right| \\
&\leq \sup_{\mathbf{B} \in B(1,\mathbf{0})} \sum_{i,j=1}^{J} |B_{ij}| \left(\left| \frac{1}{n} \sum_{a=1}^{n} k(\mathbf{x}_a, \mathbf{v}_i) k(\mathbf{x}_a, \mathbf{v}_j) - \mathbb{E}_{\mathbf{x}} \left[k(\mathbf{x}, \mathbf{v}_i) k(\mathbf{x}, \mathbf{v}_j) \right] \right| \right. \\
&\quad + \left| \frac{1}{n} \sum_{a=1}^{n} k(\mathbf{x}_a, \mathbf{v}_i) k(\mathbf{y}_a, \mathbf{v}_j) - \mathbb{E}_{\mathbf{xy}} \left[k(\mathbf{x}, \mathbf{v}_i) k(\mathbf{y}, \mathbf{v}_j) \right] \right| \\
&\quad + \left| \frac{1}{n} \sum_{a=1}^{n} k(\mathbf{y}_a, \mathbf{v}_i) k(\mathbf{x}_a, \mathbf{v}_j) - \mathbb{E}_{\mathbf{xy}} \left[k(\mathbf{y}, \mathbf{v}_i) k(\mathbf{x}, \mathbf{v}_j) \right] \right| \\
&\quad \left. + \left| \frac{1}{n} \sum_{a=1}^{n} k(\mathbf{y}_a, \mathbf{v}_i) k(\mathbf{y}_a, \mathbf{v}_j) - \mathbb{E}_{\mathbf{y}} \left[k(\mathbf{y}, \mathbf{v}_i) k(\mathbf{y}, \mathbf{v}_j) \right] \right| \right) \\
&\leq J \sup_{\mathbf{v}, \mathbf{v}' \in \mathcal{X}} \sup_{k \in \mathcal{K}} \left| \frac{1}{n} \sum_{a=1}^{n} k(\mathbf{x}_a, \mathbf{v}) k(\mathbf{x}_a, \mathbf{v}') - \mathbb{E}_{\mathbf{x}} \left[k(\mathbf{x}, \mathbf{v}) k(\mathbf{x}, \mathbf{v}') \right] \right|
\end{aligned}
$$

$$+2J \sup_{\mathbf{v},\mathbf{v}'\in\mathcal{X}} \sup_{k\in\mathcal{K}} \left| \frac{1}{n}\sum_{a=1}^{n} k(\mathbf{x}_a,\mathbf{v})k(\mathbf{y}_a,\mathbf{v}') - \mathbb{E}_{\mathbf{xy}}\left[k(\mathbf{x},\mathbf{v})k(\mathbf{y},\mathbf{v}')\right] \right|$$
$$+J \sup_{\mathbf{v},\mathbf{v}'\in\mathcal{X}} \sup_{k\in\mathcal{K}} \left| \frac{1}{n}\sum_{a=1}^{n} k(\mathbf{y}_a,\mathbf{v})k(\mathbf{y}_a,\mathbf{v}') - \mathbb{E}_{\mathbf{y}}\left[k(\mathbf{y},\mathbf{v})k(\mathbf{y},\mathbf{v}')\right] \right|$$

by exploiting that $\|\mathbf{A}\|_F = \sup_{\mathbf{B}\in B(1,\mathbf{0})} \langle \mathbf{B},\mathbf{A}\rangle_F$, and $\sum_{i,j=1}^{J} |B_{ij}| \leq J\|\mathbf{B}\|_F \leq J$ with $\mathbf{B}\in B(1,\mathbf{0})$. Using the bounds obtained for the two terms of Eq. (3.19), we get

$$\begin{aligned}
&\sup_{\mathcal{V}} \sup_{k\in\mathcal{K}} \|\mathbf{S}_n(\mathcal{V}) - \mathbf{\Sigma}(\mathcal{V})\|_F \\
&\leq \frac{8B^2J}{n-1} + 2B\sqrt{J}\frac{2n-1}{n-1} \sup_{\mathcal{V}} \sup_{k\in\mathcal{K}} \|\bar{\mathbf{z}}_n - \boldsymbol{\mu}(\mathcal{V})\|_2 \\
&+ J\left(\|P_n - P\|_{\mathcal{F}_2} + 2\|(P\times Q)_n - (P\times Q)\|_{\mathcal{F}_3} + \|Q_n - Q\|_{\mathcal{F}_2} \right). \qquad (3.21)
\end{aligned}$$

3.A.4 Bounding by Concentration and the VC Property

By combining Eqs. (3.14), (3.18) and (3.21)

$$\begin{aligned}
&\sup_{\mathcal{V}} \sup_{k} \left| \bar{\mathbf{z}}_n^\top (\mathbf{S}_n + \gamma_n I)^{-1} \bar{\mathbf{z}}_n - \boldsymbol{\mu}^\top \mathbf{\Sigma}^{-1} \boldsymbol{\mu} \right| \leq \\
&\leq \frac{\bar{c}_1}{\gamma_n}\left[\frac{8B^2J}{n-1} + 2B\sqrt{J}\frac{2n-1}{n-1}\sqrt{J}\left(\|P_n - P\|_{\mathcal{F}_1} + \|Q_n - Q\|_{\mathcal{F}_1} \right) \right. \\
&\quad \left. + J\left(\|P_n - P\|_{\mathcal{F}_2} + 2\|(P\times Q)_n - (P\times Q)\|_{\mathcal{F}_3} + \|Q_n - Q\|_{\mathcal{F}_2} \right) \right] \\
&\quad + \bar{c}_2\sqrt{J}\left(\|P_n - P\|_{\mathcal{F}_1} + \|Q_n - Q\|_{\mathcal{F}_1} \right) + \bar{c}_3\gamma_n \\
&= \left(\|P_n - P\|_{\mathcal{F}_1} + \|Q_n - Q\|_{\mathcal{F}_1} \right)\left(\frac{2}{\gamma_n}\bar{c}_1 BJ\frac{2n-1}{n-1} + \bar{c}_2\sqrt{J} \right) + \bar{c}_3\gamma_n \\
&\quad + \frac{\bar{c}_1}{\gamma_n} J\left[\|P_n - P\|_{\mathcal{F}_2} + \|Q_n - Q\|_{\mathcal{F}_2} + 2\|(P\times Q)_n - (P\times Q)\|_{\mathcal{F}_3} \right] + \frac{8}{\gamma_n}\frac{\bar{c}_1 B^2 J}{n-1} \qquad (3.22)
\end{aligned}$$

Applying Lemma 3.7 with $\frac{\delta}{5}$, we get the statement with a union bound. □

Lemma 3.7 (Concentration of the empirical process for uniformly bounded separable Carathéodory VC classes)**.** *Let $\mathcal{F}$ be*

1. *VC-subgraph class of $\mathcal{M}\to\mathbb{R}$ functions with VC index VC($\mathcal{F}$),*
2. *a uniformly bounded ($\|f\|_{L^\infty(\mathcal{M})} \leq K < \infty, \forall f\in\mathcal{F}$) separable Carathéodory family.*

Let $\mathbb{Q}$ be a probability measure, and let $\mathbb{Q}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ be the corresponding empirical measure. Then for any $\delta\in(0,1)$ with probability at least $1-\delta$

$$\|\mathbb{Q}_n - \mathbb{Q}\|_{\mathcal{F}} \leq \frac{16\sqrt{2}K}{\sqrt{n}}\left[2\sqrt{\log\left[C\times VC(\mathcal{F})(16e)^{VC(\mathcal{F})}\right]} + \frac{\sqrt{2\pi[VC(\mathcal{F})-1]}}{2} \right] + K\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}}$$

where the universal constant C is according to Lemma 3.11(iv).

Proof. Notice that $g(x_1,\ldots,x_n) = \|\mathbb{Q}_n - \mathbb{Q}\|_{\mathcal{F}}$ satisfies the bounded difference property with $b = \frac{2K}{n}$ [see Eq. (3.30)]:

$$
\begin{aligned}
&|g(\mathbf{x}_1,\ldots,\mathbf{x}_n) - g\left(\mathbf{x}_1,\ldots,\mathbf{x}_j,\mathbf{x}'_j,\mathbf{x}_{j+1},\ldots,\mathbf{x}_n\right)| \\
&\leq \left|\sup_{f\in\mathcal{F}}\left|\mathbb{Q}f - \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i)\right| - \sup_{f\in\mathcal{F}}\left|\mathbb{Q}f - \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i) + \frac{1}{n}\left[f(\mathbf{x}_j) - f(\mathbf{x}'_j)\right]\right|\right| \\
&\leq \frac{1}{n}\sup_{f\in\mathcal{F}}|f(\mathbf{x}_j) - f(\mathbf{x}'_j)| \leq \frac{1}{n}\left(\sup_{f\in\mathcal{F}}|f(\mathbf{x}_j)| + \sup_{f\in\mathcal{F}}|f(\mathbf{x}'_j)|\right) \leq \frac{2K}{n}.
\end{aligned}
$$

Hence, applying Lemma 3.12, and using symmetrization [Steinwart and Christmann, 2008] (Prop. 7.10) for the uniformly bounded separable Carathéodory $\mathcal{F}$ class, for arbitrary $\delta \in (0,1)$ with probability at least $1-\delta$

$$
\begin{aligned}
\|\mathbb{Q}_n - \mathbb{Q}\|_{\mathcal{F}} &\leq \mathbb{E}_{x_{1:n}}\|\mathbb{Q}_n - \mathbb{Q}\|_{\mathcal{F}} + K\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}} \\
&\leq 2\mathbb{E}_{x_{1:n}} R(\mathcal{F}, x_{1:n}) + K\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}}.
\end{aligned}
$$

By the Dudley entropy bound [Bousquet, 2003, Eq. (4.4)], Lemma 3.11(iv) [with $F \equiv K$, $q = 2$ $\mathbb{M} = \mathbb{Q}_n$] and the monotone decreasing property of the covering number, one arrives at

$$
\begin{aligned}
&R(\mathcal{F}, x_{1:n}) \\
&\leq \frac{8\sqrt{2}}{\sqrt{n}}\int_0^{2K}\sqrt{\log N(r,\mathcal{F},L^2(\mathcal{M},\mathbb{Q}_n))}\mathrm{d}r \\
&\leq \frac{8\sqrt{2}}{\sqrt{n}}\left[\int_0^{K}\sqrt{\log N(r,\mathcal{F},L^2(\mathcal{M},\mathbb{Q}_n))}\mathrm{d}r + K\sqrt{\log N(K,\mathcal{F},L^2(\mathcal{M},\mathbb{Q}_n))}\right] \\
&\leq \frac{8\sqrt{2}K}{\sqrt{n}}\left[\int_0^{1}\sqrt{\log N(rK,\mathcal{F},L^2(\mathcal{M},\mathbb{Q}_n))}\mathrm{d}r + \sqrt{\log N(K,\mathcal{F},L^2(\mathcal{M},\mathbb{Q}_n))}\right] \\
&\leq \frac{8\sqrt{2}K}{\sqrt{n}}\left[\int_0^{1}\sqrt{\log\left[a_1\left(\frac{1}{r}\right)^{a_2}\right]}\mathrm{d}r + \sqrt{\log(a_1)}\right] \\
&= \frac{8\sqrt{2}K}{\sqrt{n}}\left[2\sqrt{\log(a_1)} + \int_0^{1}\sqrt{a_2\log\left(\frac{1}{r}\right)}\mathrm{d}r\right] \\
&= \frac{8\sqrt{2}K}{\sqrt{n}}\left[2\sqrt{\log(a_1)} + \frac{\sqrt{\pi a_2}}{2}\right],
\end{aligned}
$$

where $a_1 := C \times VC(\mathcal{F})(16e)^{VC(\mathcal{F})}$, $a_2 := 2[VC(\mathcal{F}) - 1]$ and $\int_0^1\sqrt{\log\left(\frac{1}{r}\right)}\mathrm{d}r = \int_0^\infty t^{\frac{1}{2}}e^{-t}\mathrm{d}t = \Gamma\left(\frac{3}{2}\right) = \frac{\sqrt{\pi}}{2}$. □

Lemma 3.8 (Properties of $\mathcal{F}_i$ from $\mathcal{K}$)**.**

1. ***Uniform boundedness of $\mathcal{F}_i$-s [see Eqs. (3.11)-(3.12)]:*** *If $\mathcal{K}$ is uniformly bounded, i.e.,*

$\exists B < \infty$ such that $\sup_{k\in\mathcal{K}} \sup_{(\mathbf{x},\mathbf{y})\in\mathcal{X}^2} |k(\mathbf{x},\mathbf{y})| \leq B$; then $\mathcal{F}_1$, $\mathcal{F}_2$ and $\mathcal{F}_3$ [Eqs. (3.11)-(3.12)] are also uniformly bounded with B, B^2, B^2 constants, respectively. That is,

$$\sup_{k\in\mathcal{K},\mathbf{v}\in\mathcal{X}} |k(\mathbf{x},\mathbf{v})| \leq B,$$
$$\sup_{k\in\mathcal{K},(\mathbf{v},\mathbf{v}')\in\mathcal{X}^2} |k(\mathbf{x},\mathbf{v})k(\mathbf{x},\mathbf{v}')| \leq B^2,$$
$$\sup_{k\in\mathcal{K},(\mathbf{v},\mathbf{v}')\in\mathcal{X}^2} |k(\mathbf{x},\mathbf{v})k(\mathbf{y},\mathbf{v}')| \leq B^2.$$

2. ***Separability of $\mathcal{F}_i$:*** *since $\mathcal{F}_1$, $\mathcal{F}_2$ and $\mathcal{F}_3$ is parameterized by $\Theta = \mathcal{K}\times\mathcal{X}$, $\mathcal{K}\times\mathcal{X}^2$, $\mathcal{K}\times\mathcal{X}^2$, separability of $\mathcal{K}$ implies that of Θ.*

3. ***Measurability of $\mathcal{F}_i$:*** *$\forall k \in \mathcal{K}$ is measurable, then the elements of $\mathcal{F}_i$ ($i = 1,2,3$) are also measurable.* □

3.A.5 Finite VC Index of the Gaussian Kernel Class

Recall that Theorem 3.6 requires that the kernel class $\mathcal{K}$ consist of measurable kernels, and be uniformly bounded (condition 1). Further, it also requires that $\mathcal{F}_i$-s (associated with $\mathcal{K}$) be a VC-subgraph, separable Carathéodory families (condition 3). In this section, we show that the kernel class consisting of all the isotropic Gaussian kernels satisfies these two conditions. The VC property will be a direct consequence of the VC indices of finite-dimensional function classes and preservation theorems (see Lemma 3.11); for a nice example application see Srebro and Ben-David [2006] (Section 5) who study the pseudo-dimension of $(\mathbf{x},\mathbf{y}) \mapsto k(\mathbf{x},\mathbf{y})$ kernel classes, for different Gaussian families. We take the Gaussian class and use the preservation trick to bound the VC index of the associated $\mathcal{F}_i$-s.

Lemma 3.9 ($\mathcal{F}_i$-s are VC-subgraph and uniformly bounded separable Carathéodory families for the isotropic Gaussian kernel)**.** *Let*

$$\mathcal{K} = \left\{ k_\sigma : (\mathbf{x},\mathbf{y}) \in \mathcal{X}\times\mathcal{X} \subseteq \mathbb{R}^d \times \mathbb{R}^d \mapsto e^{-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2\sigma^2}} : \sigma > 0 \right\}.$$

Then, the $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_3$ classes [see Eqs. (3.11)-(3.12)] associated to $\mathcal{K}$ are

- *VC-subgraphs with indices $VC(\mathcal{F}_1) \leq d+4$, $VC(\mathcal{F}_2) \leq d+4$, $VC(\mathcal{F}_3) \leq 2d+4$, and*
- *uniformly bounded separable Carathéodory families, with $\|f\|_{L^\infty(\mathcal{M})} \leq 1$ for all $f \in \{\mathcal{F}_1,\mathcal{F}_2,\mathcal{F}_3\}$.*[5]

Proof. **VC subgraph property:**

[5] $\mathcal{M} = \mathcal{X}$ for $\mathcal{F}_1$ and $\mathcal{F}_2$, and $\mathcal{M} = \mathcal{X}^2$ in case of $\mathcal{F}_3$.

$\mathcal{F}_1$: Consider the function class

$$\mathcal{G} = \left\{ \mathbf{x} \mapsto \frac{\|\mathbf{x}-\mathbf{v}\|_2^2}{2\sigma^2} = \frac{1}{2\sigma^2}\left(\|\mathbf{x}\|_2^2 - 2\langle \mathbf{x}, \mathbf{v}\rangle_2 + \|\mathbf{v}\|_2^2\right) : \sigma > 0, \mathbf{v} \in \mathcal{X} \right\} \subseteq L^0(\mathbb{R}^d),$$

where $\mathcal{G} \subseteq \tilde{\mathcal{G}} := span\left(\mathbf{x} \mapsto \|\mathbf{x}\|_2^2, \{\mathbf{x} \mapsto x_i\}_{i=1}^d, \mathbf{x} \mapsto 1\right)$ is a vector space, and $dim(\mathcal{G}) \leq d+2$. Thus by Lemma 3.11(i)-(ii), $\mathcal{G}$ is VC with $VC(\mathcal{G}) \leq d+4$; applying Lemma 3.11(iii) with $\phi(z) = e^{-z}$, $\mathcal{F}_1 = \phi \circ \mathcal{G}$ is also VC with index $VC(\mathcal{F}_1) \leq d+4$.

$\mathcal{F}_2$: Since

$$\mathcal{F}_2 = \left\{ \mathbf{x} \mapsto k(\mathbf{x}, \mathbf{v})k(\mathbf{x}, \mathbf{v}') = e^{-\frac{\|\mathbf{x}-\mathbf{v}\|_2^2 + \|\mathbf{x}-\mathbf{v}'\|_2^2}{2\sigma^2}} : \sigma > 0, \mathbf{v} \in \mathcal{X}, \mathbf{v}' \in \mathcal{X} \right\},$$

and $\left\{ \mathbf{x} \mapsto \frac{\|\mathbf{x}-\mathbf{v}\|_2^2 + \|\mathbf{x}-\mathbf{v}'\|_2^2}{2\sigma^2} : \sigma > 0, \mathbf{v} \in \mathcal{X}, \mathbf{v}' \in \mathcal{X} \right\}$ is a subset of $S = span\big(\mathbf{x} \mapsto \|\mathbf{x}\|_2^2, \{\mathbf{x} \mapsto x_i\}_{i=1}^d, \mathbf{x} \mapsto 1\big)$, it follows that $VC(\mathcal{F}_2) \leq d+4$.

$\mathcal{F}_3$: Since

$$\mathcal{F}_3 = \left\{ (\mathbf{x}, \mathbf{y}) \mapsto k(\mathbf{x}, \mathbf{v})k(\mathbf{y}, \mathbf{v}') = e^{-\frac{\|\mathbf{x}-\mathbf{v}\|^2 + \|\mathbf{y}-\mathbf{v}'\|_2^2}{2\sigma^2}} = e^{-\frac{\|[\mathbf{x};\mathbf{y}]-[\mathbf{v};\mathbf{v}']\|_2^2}{2\sigma^2}} : \sigma > 0, \mathbf{v} \in \mathbb{R}^d, \mathbf{v}' \in \mathbb{R}^d \right\},$$

from the result on $\mathcal{F}_1$ we get that $VC(\mathcal{F}_3) \leq 2d+4$.

Uniformly bounded, separable Carathéodory family The result follows from Lemma 3.8 by noting that $|k(\mathbf{x}, \mathbf{y})| \leq 1 =: B$, $(\mathbf{x}, \mathbf{y}) \mapsto e^{-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2\sigma^2}}$ is continuous ($\forall \sigma > 0$), $\mathbb{R}^+$ is separable, and the $(\sigma, \mathbf{v}) \mapsto e^{-\frac{\|\mathbf{x}-\mathbf{v}\|_2^2}{2\sigma^2}}$, $(\sigma, \mathbf{v}, \mathbf{v}') \mapsto e^{-\frac{\|\mathbf{x}-\mathbf{v}\|_2^2}{2\sigma^2}} e^{-\frac{\|\mathbf{x}-\mathbf{v}'\|_2^2}{2\sigma^2}}$, $(\sigma, \mathbf{v}, \mathbf{v}') \mapsto e^{-\frac{\|\mathbf{x}-\mathbf{v}\|_2^2}{2\sigma^2}} e^{-\frac{\|\mathbf{y}-\mathbf{v}'\|_2^2}{2\sigma^2}}$ mappings are continuous ($\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$). □

We note that to also satisfy condition 2 of Theorem 3.6 (i.e., $\tilde{c} := \sup_{\mathcal{V} \in \mathbb{V}} \sup_{k \in \mathcal{K}} \|\boldsymbol{\Sigma}^{-1}\|_F < \infty$), $\mathcal{K}$ requires further restriction (see Section 3.4.2).

3.B Proof: A Lower Bound on the Test Power

Recall Proposition 3.5:

Proposition 3.10 (Lower bound on the test power). *Define $\boldsymbol{\mu} := \mathbb{E}_{\mathbf{xy}}[\mathbf{z}_1]$, $\boldsymbol{\Sigma} := \mathbb{E}_{\mathbf{xy}}[(\mathbf{z}_1 - \boldsymbol{\mu})(\mathbf{z}_1 - \boldsymbol{\mu})^\top]$, and $\lambda_n := n\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$ (the population counterpart of $\hat{\lambda}_n$). Let $\mathbb{V}$ be a collection in which each element is a set of J test locations.*

- *For the ME test, let $\mathcal{K}$ be a uniformly bounded (i.e., there exists $B < \infty$ such that $\sup_{k \in \mathcal{K}} \sup_{(\mathbf{x},\mathbf{y}) \in \mathcal{X}^2} |k(\mathbf{x}, \mathbf{y})| \leq B$) family of $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ measurable kernels.*
- *For the SCF test, let $\mathcal{K}$ be a class of translation-invariant kernels such that $\mathcal{L} = \{\mathbf{x} \mapsto \hat{k}(\mathbf{x}) : k \in \mathcal{K}\}$ is uniformly bounded (i.e., $\exists B < \infty$ such that $\sup_{f \in \mathcal{L}} \sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x})| \leq B$) family, where $\hat{k}(\mathbf{x}) := \int e^{-i\mathbf{w}^\top \mathbf{x}} k(\mathbf{w}) \, \mathrm{d}\mathbf{w}$.*

Assume that $\tilde{c} := \sup_{\mathcal{V}\in\mathbb{V},k\in\mathcal{K}} \|\boldsymbol{\Sigma}^{-1}\|_F < \infty$. *Then, for any* $\mathcal{V} \in \mathbb{V}$, *for large* n, *the test power* $\mathbb{P}\left(\hat{\lambda}_n \geq T_\alpha\right)$ *of both tests satisfies* $\mathbb{P}\left(\hat{\lambda}_n \geq T_\alpha\right) \geq L(\lambda_n)$ *where*

$$L(\lambda_n) := 1 - 2e^{-\frac{(\lambda_n - T_\alpha)^2}{3^2 \cdot 8B^2\bar{c}_2^2 J' n}} - 2e^{-\frac{(\gamma_n(\lambda_n - T_\alpha)(n-1) - 24B^2\bar{c}_1 J' n)^2}{3^2 \cdot 32B^4\bar{c}_1^2 J'^2 n(2n-1)^2}} - 2e^{-\frac{\left((\lambda_n - T_\alpha)/3 - \bar{c}_3 n\gamma_n\right)^2 \gamma_n^2}{32B^4 J'^2 \bar{c}_1^2 n}},$$

$\bar{c}_1 := 4B^2 J'\sqrt{J}\tilde{c}, \bar{c}_2 := 4B\sqrt{J'}\tilde{c}$, *and* $\bar{c}_3 := 4B^2 J'\tilde{c}^2$. *For the ME test,* $J' = J$. *For the SCF test,* $J' = 2J$. *For large* n, $L(\lambda_n)$ *is increasing in* λ_n.

Proof

We first consider the case of the ME test. By (3.14), we have

$$|\hat{\lambda}_n - \lambda_n| \leq \frac{\bar{c}_1 n}{\gamma_n}\|\boldsymbol{\Sigma} - \mathbf{S}_n\|_F + \bar{c}_2 n\|\bar{\mathbf{z}}_n - \boldsymbol{\mu}\|_2 + \bar{c}_3 n\gamma_n. \tag{3.23}$$

We will bound each of the three terms in (3.23).

Bounding $\|\bar{\mathbf{z}}_n - \boldsymbol{\mu}\|_2$ (Second Term in (3.23))

Let

$$g(\mathbf{x}, \mathbf{y}, \mathbf{v}) := k(\mathbf{x}, \mathbf{v}) - k(\mathbf{y}, \mathbf{v}). \tag{3.24}$$

Define $\mathbf{v}^* := \arg\max_{\mathbf{v}\in\{\mathbf{v}_1,\ldots,\mathbf{v}_J\}} \left|\frac{1}{n}\sum_{i=1}^n g(\mathbf{x}_i, \mathbf{y}_i, \mathbf{v}) - \mathbb{E}_{\mathbf{xy}}\left[g(\mathbf{x}, \mathbf{y}, \mathbf{v})\right]\right|$ and $G_i := g(\mathbf{x}_i, \mathbf{y}_i, \mathbf{v}^*)$.

$$\begin{aligned}
\|\bar{\mathbf{z}}_n - \boldsymbol{\mu}\|_2 &= \sup_{\mathbf{b}\in B(1,\mathbf{0})} \langle \mathbf{b}, \bar{\mathbf{z}}_n - \boldsymbol{\mu}\rangle_2 \\
&\leq \sup_{\mathbf{b}\in B(1,\mathbf{0})} \sum_{j=1}^J |b_j| \left|\frac{1}{n}\sum_{i=1}^n \left[k(\mathbf{x}_i, \mathbf{v}_j) - k(\mathbf{y}_i, \mathbf{v}_j)\right] - \mathbb{E}_{\mathbf{xy}}\left[k(\mathbf{x}, \mathbf{v}_j) - k(\mathbf{y}, \mathbf{v}_j)\right]\right| \\
&= \sup_{\mathbf{b}\in B(1,\mathbf{0})} \sum_{j=1}^J |b_j| \left|\frac{1}{n}\sum_{i=1}^n g(\mathbf{x}_i, \mathbf{y}_i, \mathbf{v}_j) - \mathbb{E}_{\mathbf{xy}}\left[g(\mathbf{x}, \mathbf{y}, \mathbf{v}_j)\right]\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^n G_i - \mathbb{E}_{\mathbf{xy}}[G_1]\right| \sup_{\mathbf{b}\in B(1,\mathbf{0})} \sum_{j=1}^J |b_j| \\
&\leq \sqrt{J}\left|\frac{1}{n}\sum_{i=1}^n G_i - \mathbb{E}_{\mathbf{xy}}[G_1]\right| \sup_{\mathbf{b}\in B(1,\mathbf{0})} \|\mathbf{b}\|_2 \\
&= \sqrt{J}\left|\frac{1}{n}\sum_{i=1}^n G_i - \mathbb{E}_{\mathbf{xy}}[G_1]\right|,
\end{aligned}$$

where we used the fact that $\|\mathbf{b}\|_1 \leq \sqrt{J}\|\mathbf{b}\|_2$. It can be seen that $-2B \leq G_i \leq 2B$ because

$$G_i = k(\mathbf{x}_i, \mathbf{v}^*) - k(\mathbf{y}_i, \mathbf{v}^*) \leq |k(\mathbf{x}_i, \mathbf{v}^*)| + |k(\mathbf{y}_i, \mathbf{v}^*)| \leq 2B.$$

Using Hoeffding's inequality (Lemma 3.13) to bound $\left|\frac{1}{n}\sum_{i=1}^n G_i - \mathbb{E}_{\mathbf{xy}}[G_1]\right|$, we have

$$\mathbb{P}\left(n\bar{c}_2\|\bar{\mathbf{z}}_n - \boldsymbol{\mu}\|_2 \leq \alpha\right) \leq 1 - 2\exp\left(-\frac{\alpha^2}{8B^2\bar{c}_2^2 Jn}\right). \tag{3.25}$$

Bounding First ($\|\boldsymbol{\Sigma} - \mathbf{S}_n\|_F$) and Third Terms in (3.23)

Let $\eta(\mathbf{v}_i, \mathbf{v}_j) := \left|\frac{1}{n}\sum_{a=1}^n g(\mathbf{x}_a, \mathbf{y}_a, \mathbf{v}_i)g(\mathbf{x}_a, \mathbf{y}_a, \mathbf{v}_j) - \mathbb{E}_{\mathbf{xy}}\left[g(\mathbf{x}, \mathbf{y}, \mathbf{v}_i)g(\mathbf{x}, \mathbf{y}, \mathbf{v}_j)\right]\right|$. Define $(\mathbf{v}_1^*, \mathbf{v}_2^*) = \arg\max_{(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}) \in \{(\mathbf{v}_i, \mathbf{v}_j)\}_{i,j=1}^J} \eta(\mathbf{v}^{(1)}, \mathbf{v}^{(2)})$. Define $H_i := g(\mathbf{x}_i, \mathbf{y}_i, \mathbf{v}_1^*)g(\mathbf{x}_i, \mathbf{y}_i, \mathbf{v}_2^*)$. By (3.19), we have

$$\begin{aligned}
\|\mathbf{S}_n - \boldsymbol{\Sigma}\|_F &\leq (*_1) + (*_2), \\
(*_1) &= \left\| \frac{1}{n}\sum_{a=1}^n \mathbf{z}_a \mathbf{z}_a^\top - \mathbb{E}_{\mathbf{xy}}[\mathbf{z}_1 \mathbf{z}_1^\top] \right\|_F, \\
(*_2) &= \frac{8B^2 J}{n-1} + 2B_k\sqrt{J}\frac{2n-1}{n-1}\|\bar{\mathbf{z}}_n - \boldsymbol{\mu}\|_2.
\end{aligned}$$

We can upper bound $(*_2)$ by applying Hoeffding's inequality to bound $\|\bar{\mathbf{z}}_n - \boldsymbol{\mu}\|_2$ giving

$$\mathbb{P}\left(\frac{\bar{c}_1 n}{\gamma_n}(*_2) \leq \alpha\right) \geq 1 - 2\exp\left(-\frac{(\alpha\gamma_n - \alpha\gamma_n n + 8B^2\bar{c}_1 J n)^2}{32B^4\bar{c}_1^2 J^2 n(2n-1)^2}\right). \tag{3.26}$$

We can upper bound $(*_1)$ with

$$\begin{aligned}
(*_1) &= \sup_{\mathbf{B} \in B(1,\mathbf{0})} \left\langle \mathbf{B}, \frac{1}{n}\sum_{a=1}^n \mathbf{z}_a \mathbf{z}_a^\top - \mathbb{E}_{\mathbf{xy}}[\mathbf{z}_1 \mathbf{z}_1^\top] \right\rangle_F \\
&\leq \sup_{\mathbf{B} \in B(1,\mathbf{0})} \sum_{i=1}^J \sum_{j=1}^J |B_{ij}| \left| \frac{1}{n}\sum_{a=1}^n g(\mathbf{x}_a, \mathbf{y}_a, \mathbf{v}_i)g(\mathbf{x}_a, \mathbf{y}_a, \mathbf{v}_j) - \mathbb{E}_{\mathbf{xy}}\left[g(\mathbf{x}, \mathbf{y}, \mathbf{v}_i)g(\mathbf{x}, \mathbf{y}, \mathbf{v}_j)\right] \right| \\
&\leq \left| \frac{1}{n}\sum_{a=1}^n H_a - \mathbb{E}_{\mathbf{xy}}[H_1] \right| \sup_{\mathbf{B} \in B(1,\mathbf{0})} \sum_{i=1}^J \sum_{j=1}^J |B_{ij}| \\
&\leq J \left| \frac{1}{n}\sum_{a=1}^n H_a - \mathbb{E}_{\mathbf{xy}}[H_1] \right| \sup_{\mathbf{B} \in B(1,\mathbf{0})} \|\mathbf{B}\|_F = J \left| \frac{1}{n}\sum_{a=1}^n H_a - \mathbb{E}_{\mathbf{xy}}[H_1] \right|,
\end{aligned}$$

where we used the fact that $\sum_{i=1}^J \sum_{j=1}^J |B_{ij}| \leq J\|\mathbf{B}\|_F$. It can be seen that $-4B^2 \leq H_a \leq 4B^2$. Using Hoeffding's inequality (Lemma 3.13) to bound $\left|\frac{1}{n}\sum_{a=1}^n H_a - \mathbb{E}_{\mathbf{xy}}[H_1]\right|$, we have

$$\mathbb{P}\left(\frac{\bar{c}_1 n}{\gamma_n}(*_1) \leq \alpha\right) \geq 1 - 2\exp\left(-\frac{\alpha^2\gamma_n^2}{32B^4 J^2 \bar{c}_1^2 n}\right), \tag{3.27}$$

implying that

$$\mathbb{P}\left(\frac{\bar{c}_1 n}{\gamma_n}(*_1) + \bar{c}_3 n\gamma_n \leq \alpha\right) \geq 1 - 2\exp\left(-\frac{(\alpha - \bar{c}_3 n\gamma_n)^2 \gamma_n^2}{32B^4 J^2 \bar{c}_1^2 n}\right). \tag{3.28}$$

Applying a union bound on (3.25), (3.26), and (3.28) with $t = \alpha/3$, we can conclude that

$$\mathbb{P}\left(\left|\hat{\lambda}_n - \lambda_n\right| \leq t\right) \geq \mathbb{P}\left(\frac{\bar{c}_1 n}{\gamma_n}\|\boldsymbol{\Sigma} - \mathbf{S}_n\|_F + \bar{c}_2 n\|\bar{\mathbf{z}}_n - \boldsymbol{\mu}\|_2 + \bar{c}_3 n\gamma_n \leq t\right)$$

$$\geq 1 - 2\exp\left(-\frac{t^2}{3^2 \cdot 8B^2\bar{c}_2^2 Jn}\right) - 2\exp\left(-\frac{(t\gamma_n n - t\gamma_n - 24B^2\bar{c}_1 Jn)^2}{3^2 \cdot 32B^4\bar{c}_1^2 J^2 n(2n-1)^2}\right) - 2\exp\left(-\frac{(t/3 - \bar{c}_3 n\gamma_n)^2\,\gamma_n^2}{32B^4 J^2\bar{c}_1^2 n}\right).$$

A rearrangement yields

$$\mathbb{P}\left(\hat{\lambda}_n \geq T_\alpha\right) \geq 1 - 2e^{-\frac{(\lambda_n - T_\alpha)^2}{3^2 \cdot 8B^2\bar{c}_2^2 Jn}} - 2e^{-\frac{(\gamma_n(\lambda_n - T_\alpha)(n-1) - 24B^2\bar{c}_1 Jn)^2}{3^2 \cdot 32B^4\bar{c}_1^2 J^2 n(2n-1)^2}} - 2e^{-\frac{\left((\lambda_n - T_\alpha)/3 - \bar{c}_3 n\gamma_n\right)^2\gamma_n^2}{32B^4 J^2\bar{c}_1^2 n}}.$$

□

Proof of the Lower Bound on the SCF Test Power

The proof follows the same structure as in the proof of the ME test lower bound given in Section 3.B. We only need to redefine g in (3.24) to be one of the following g_s and g_c. Consider $g_s(\mathbf{x}, \mathbf{y}, \mathbf{v}) := f(\mathbf{x})\sin(\mathbf{x}^\top \mathbf{v}) - f(\mathbf{y})\sin(\mathbf{y}^\top \mathbf{v})$ and $g_c(\mathbf{x}, \mathbf{y}, \mathbf{v}) := f(\mathbf{x})\cos(\mathbf{x}^\top \mathbf{v}) - f(\mathbf{y})\cos(\mathbf{y}^\top \mathbf{v})$ for an arbitrary $f \in \mathcal{L}$. It can be seen that both g_s and g_c are bounded by $2B$:

$$\begin{aligned} g_s(\mathbf{x}, \mathbf{y}, \mathbf{v}) &:= f(\mathbf{x})\sin(\mathbf{x}^\top \mathbf{v}) - f(\mathbf{y})\sin(\mathbf{y}^\top \mathbf{v}) \\ &\leq |f(\mathbf{x})\sin(\mathbf{x}^\top \mathbf{v})| + |f(\mathbf{y})\sin(\mathbf{y}^\top \mathbf{v})| \\ &\leq 2B. \end{aligned}$$

Derivation for g_c is identical. For $g \in \{g_s, g_c\}$, these bounds imply that $-2B \leq g(\mathbf{x}, \mathbf{y}, \mathbf{v}) \leq 2B$ for any $\mathbf{x}, \mathbf{y}, \mathbf{v}$. Since $|g(\mathbf{x}, \mathbf{y}, \mathbf{v})| \leq 2B$ which is the same as in Section 3.B, the bound follows immediate where we only need to replace J with $J' = 2J$. This replacement follows from the fact that there are J features for sin and J features for cos, totaling $2J$. □

3.C External Lemmas

In this section we detail some external lemmas used in our proofs.

Lemma 3.11 (Properties of VC classes, see page 141, 146-147 in van der Vaart [2000] and page 160-161 in Kosorok [2008])**.**

(i) Monotonicity: $\mathcal{G} \subseteq \tilde{\mathcal{G}} \subseteq L^0(\mathcal{M}) \Rightarrow VC(\mathcal{G}) \leq VC(\tilde{\mathcal{G}})$.

(ii) Finite-dimensional vector space: if $\mathcal{G}$ *is a finite-dimensional vector space of measurable functions, then* $VC(\mathcal{G}) \leq dim(\mathcal{G}) + 2$.

(iii) Composition with monotone function: If $\mathcal{G}$ *is VC and* $\phi : \mathbb{R} \to \mathbb{R}$ *is monotone, then for* $\phi \circ \mathcal{G} := \{\phi \circ g : g \in \mathcal{G}\}$, $VC(\phi \circ \mathcal{G}) \leq VC(\mathcal{G})$.

(iv) The r-covering number of a VC class grows only polynomially in $\frac{1}{r}$*: Let* $\mathcal{F}$ *be VC on the domain* $\mathcal{M}$ *with measurable envelope* F ($|f(m)| \leq F(m), \forall m \in \mathcal{M}, f \in \mathcal{F}$)*. Then for*

any $q \geq 1$ and $\mathbb{M}$ probability measure for which $\|F\|_{L^q(\mathcal{M},\mathbb{M})} > 0$

$$N\left(r\|F\|_{L^q(\mathcal{M},\mathbb{M})}, \mathcal{F}, L^q(\mathcal{M},\mathbb{M})\right) \leq C \times VC(\mathcal{F})(16e)^{VC(\mathcal{F})}\left(\frac{1}{r}\right)^{q[VC(\mathcal{F})-1]} \tag{3.29}$$

for any $r \in (0,1)$ with a universal constant C.

Lemma 3.12 (McDiarmid's inequality)**.** *Let $X_1, \ldots, X_n \in \mathcal{M}$ be independent random variables and let $g : \mathcal{M}^n \to \mathbb{R}$ be a function such that the*

$$\sup_{\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{x}_j' \in \mathcal{M}} \left|g(\mathbf{x}_1,\ldots,\mathbf{x}_n) - g\left(\mathbf{x}_1,\ldots,\mathbf{x}_j,\mathbf{x}_j',\mathbf{x}_{j+1},\ldots,\mathbf{x}_n\right)\right| \leq b \tag{3.30}$$

bounded difference property holds. Then for arbitrary $\delta \in (0,1)$

$$\mathbb{P}\left(g(X_1,\ldots,X_n) \leq \mathbb{E}[g(X_1,\ldots,X_n)] + b\sqrt{\frac{\log\left(\frac{1}{\delta}\right)n}{2}}\right) \geq 1-\delta.$$

Lemma 3.13 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables with $\mathbb{P}(a \leq X_i \leq b) = 1$. Let $\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$. Then,*

$$\mathbb{P}\left(\left|\overline{X} - \mathbb{E}[\overline{X}]\right| \leq t\right) \geq 1 - 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right).$$

Equivalently, for any $\delta \in (0,1)$, with probability at least $1-\delta$, it holds that

$$\left|\overline{X} - \mathbb{E}[\overline{X}]\right| \leq \frac{b-a}{\sqrt{2n}}\sqrt{\log(2/\delta)}.$$

Chapter 4

Informative Features for Dependence Detection

Summary A new computationally efficient dependence measure, and an adaptive statistical test of independence, are proposed. The dependence measure is the difference between analytic embeddings of the joint distribution and the product of the marginals, evaluated at a finite set of locations (features). These features are chosen so as to maximize a lower bound on the test power, resulting in a test that is data-efficient, and that runs in linear time (with respect to the sample size n). The optimized features can be interpreted as evidence to reject the null hypothesis, indicating regions in the joint domain where the joint distribution and the product of the marginals differ most. Consistency of the independence test is established, for an appropriate choice of features. In real-world benchmarks, independence tests using the optimized features perform comparably to the state-of-the-art quadratic-time HSIC test, and outperform competing $\mathcal{O}(n)$ and $\mathcal{O}(n \log n)$ tests.

4.1 Introduction

We consider the design of adaptive, nonparametric statistical tests of dependence: that is, tests of whether a joint distribution P_{xy} factorizes into the product of marginals $P_x P_y$ with the null hypothesis that $H_0 : X$ and Y are independent. While classical tests of dependence, such as Pearson's correlation and Kendall's τ, are able to detect monotonic relations between univariate variables, more modern tests can address complex interactions, for instance changes in variance of X with the value of Y. Key to many recent tests is to examine covariance or correlation between data features. These interactions become significantly harder to detect, and the features are more difficult to design, when the data reside in high dimensions.

A basic nonlinear dependence measure is the Hilbert-Schmidt Independence Criterion (HSIC), which is the Hilbert-Schmidt norm of the covariance operator between feature mappings of the random variables [Gretton et al., 2005a, 2008]. Each random variable X and Y is mapped to a respective reproducing kernel Hilbert space $\mathcal{H}_k$ and $\mathcal{H}_l$. For sufficiently rich mappings, the covariance operator norm is zero if and

only if the variables are independent. A second basic nonlinear dependence measure is the smoothed difference between the characteristic function of the joint distribution, and that of the product of marginals. When a particular smoothing function is used, the statistic corresponds to the covariance between distances of X and Y variable pairs [Feuerverger, 1993, Székely and Rizzo, 2009, Székely et al., 2007], yielding a simple test statistic based on pairwise distances. It has been shown by Sejdinovic et al. [2013] that the distance covariance (and its generalization to semi-metrics) is an instance of HSIC for an appropriate choice of kernels. A disadvantage of these feature covariance statistics, however, is that they require quadratic time to compute (besides in the special case of the distance covariance with univariate real-valued variables, where Huo and Székely [2016] achieve an $\mathcal{O}(n \log n)$ cost). Moreover, the feature covariance statistics have intractable null distributions, and either a permutation approach or the solution of an expensive eigenvalue problem [e.g. Zhang et al., 2011] is required for consistent estimation of the quantiles. Several approaches were proposed by Zhang et al. [2017] to obtain faster tests along the lines of HSIC. These include computing HSIC on finite-dimensional feature mappings chosen as random Fourier features (RFFs) [Rahimi and Recht, 2007], a block-averaged statistic, and a Nyström approximation to the statistic. Key to each of these approaches is a more efficient computation of the statistic and its threshold under the null distribution: for RFFs, the null distribution is a finite weighted sum of χ^2 variables; for the block-averaged statistic, the null distribution is asymptotically normal; for Nyström, either a permutation approach is employed, or the spectrum of the Nyström approximation to the kernel matrix is used in approximating the null distribution. Each of these methods costs significantly less than the $\mathcal{O}(n^2)$ cost of the full HSIC (the cost is linear in n, but also depends quadratically on the number of features retained). A potential disadvantage of the Nyström and Fourier approaches is that the features are not optimized to maximize test power, but are chosen randomly. The test consistency of the Nyström approximation is also not guaranteed. The block statistic performs worse than both, due to the large variance of the statistic under the null (which can be mitigated by observing more data).

In addition to feature covariances, correlation measures have also been developed in infinite dimensional feature spaces: in particular, Bach and Jordan [2002], Fukumizu et al. [2008] proposed statistics on the correlation operator in a reproducing kernel Hilbert space. While convergence has been established for certain of these statistics, their computational cost is high at $\mathcal{O}(n^3)$, and test thresholds have relied on permutation. A number of much faster approaches to testing based on feature correlations have been proposed, however. For instance, Dauxois and Nkiet [1998] compute statistics of the correlation between finite sets of basis functions, chosen for instance to be step functions or low order B-splines. The cost of this approach is $\mathcal{O}(n)$. This idea was extended by Lopez-Paz et al. [2013], who computed the canonical correlation between finite sets of basis functions chosen as random Fourier features; in addition, they performed a copula transform on the inputs, with a total cost of

$\mathcal{O}(n \log n)$. Finally, space partitioning approaches have also been proposed, based on statistics such as the KL divergence, however these apply only to univariate variables [Heller et al., 2016], or to multivariate variables of low dimension [Gretton and Györfi, 2010] (that said, these tests have other advantages of theoretical interest, notably distribution-independent test thresholds).

The approach we take is most closely related to HSIC on a finite set of features. Our simplest test statistic, the Finite Set Independence Criterion (FSIC), is an average of covariances of analytic functions (i.e., features) defined on each of X and Y. A normalized version of the statistic (NFSIC) yields a distribution-independent asymptotic test threshold. We show that our test is consistent, despite a finite number of analytic features being used, via a generalization of arguments in Chwialkowski et al. [2015]. As in recent work on two-sample testing by Jitkrittum et al. [2016], our test is *adaptive* in the sense that we choose our features on a held-out validation set to optimize a lower bound on the test power. The design of features for independence testing turns out to be quite different to the case of two-sample testing, however: the task is to find correlated feature *pairs* on the respective marginal domains, rather than attempting to find a single, high-dimensional feature representation on the *tensor product* of the marginals, as we would need to do if we were comparing distributions P_{xy} and Q_{xy}. While the use of coupled feature pairs on the marginals entails a smaller feature space dimension, it introduces significant complications in the proof of the lower bound, compared with the two-sample case. This bound converges as more validation samples are observed; since correlated feature pairs are used, a different approach is required to the analogous two-sample result of Jitkrittum et al. [2016] (described in Chapter 3). We demonstrate the performance of our tests on several challenging artificial and real-world datasets, including detection of dependence between music and its year of appearance, and between videos and captions. In these experiments, we outperform competing linear and $\mathcal{O}(n \log n)$ time tests.

4.2 New Statistic: The Finite Set Independence Criterion (FSIC)

We introduce two test statistics: first, the Finite Set Independence Criterion (FSIC), which builds on the principle that dependence can be measured in terms of the covariance between data features. Next, we propose a normalized version of this statistic (NFSIC), with a simpler asymptotic distribution when $P_{xy} = P_x P_y$. We show how to select features for the latter statistic to maximize a lower bound on the power of its corresponding statistical test.

We begin by recalling the Hilbert-Schmidt Independence Criterion (HSIC) as proposed in Gretton et al. [2005a], since our unnormalized statistic is built along similar lines. Consider two random variables $X \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $Y \in \mathcal{Y} \subseteq \mathbb{R}^{d_y}$. Denote by P_{xy} the joint distribution between X and Y; P_x and P_y are the marginal distributions of X and Y. Let $\otimes$ denote the tensor product, such that $(a \otimes b)\, c = a \langle b, c \rangle$. Assume that $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ are positive definite kernels associated with

reproducing kernel Hilbert spaces (RKHS) $\mathcal{H}_k$ and $\mathcal{H}_l$, respectively. Let $\|\cdot\|_{HS}$ be the norm on the space of $\mathcal{H}_l \to \mathcal{H}_k$ Hilbert-Schmidt operators. Then, HSIC between X and Y is defined as

$$\begin{aligned}\mathrm{HSIC}(X,Y) &= \left\|\mu_{xy} - \mu_x \otimes \mu_y\right\|^2_{\mathrm{HS}} \\ &= \mathbb{E}_{(\mathbf{x},\mathbf{y}),(\mathbf{x}',\mathbf{y}')}\left[k(\mathbf{x},\mathbf{x}')l(\mathbf{y},\mathbf{y}')\right] \\ &\quad + \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{x}'}[k(\mathbf{x},\mathbf{x}')]\mathbb{E}_{\mathbf{y}}\mathbb{E}_{\mathbf{y}'}[l(\mathbf{y},\mathbf{y}')] \\ &\quad - 2\mathbb{E}_{(\mathbf{x},\mathbf{y})}\left[\mathbb{E}_{\mathbf{x}'}[k(\mathbf{x},\mathbf{x}')]\mathbb{E}_{\mathbf{y}'}[l(\mathbf{y},\mathbf{y}')]\right], \end{aligned} \tag{4.1}$$

where $\mathbb{E}_{\mathbf{x}} := \mathbb{E}_{\mathbf{x}\sim P_x}$, $\mathbb{E}_{\mathbf{y}} := \mathbb{E}_{\mathbf{y}\sim P_y}$, $\mathbb{E}_{\mathbf{xy}} := \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P_{xy}}$, and $\mathbf{x}'$ is an independent copy of $\mathbf{x}$. The mean embedding of P_{xy} belongs to the space of Hilbert-Schmidt operators from $\mathcal{H}_l$ to $\mathcal{H}_k$, $\mu_{xy} := \int_{\mathcal{X}\times\mathcal{Y}} k(\mathbf{x},\cdot)\otimes l(\mathbf{y},\cdot)\,\mathrm{d}P_{xy}(\mathbf{x},\mathbf{y}) \in \mathrm{HS}(\mathcal{H}_l,\mathcal{H}_k)$, and the marginal mean embeddings are $\mu_x := \int_{\mathcal{X}} k(\mathbf{x},\cdot)\,\mathrm{d}P_x(\mathbf{x}) \in \mathcal{H}_k$ and $\mu_y := \int_{\mathcal{Y}} l(\mathbf{y},\cdot)\,\mathrm{d}P_y(\mathbf{y}) \in \mathcal{H}_l$ [Smola et al., 2007]. Gretton et al. [2005a, Theorem 4] show that if the kernels k and l are universal [Steinwart and Christmann, 2008] on compact domains $\mathcal{X}$ and $\mathcal{Y}$, then $\mathrm{HSIC}(X,Y) = 0$ if and only if X and Y are independent. Alternatively, Gretton [2015] shows that it is sufficient for each of k and l to be characteristic to their respective domains (meaning that distribution embeddings are injective in each marginal domain: see Sriperumbudur et al. [2010]). Given a joint sample $\mathsf{Z}_n = \{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^n \sim P_{xy}$, an empirical estimator of HSIC can be computed in $\mathcal{O}(n^2)$ time by replacing the population expectations in (4.1) with their corresponding empirical expectations based on Z_n.

Proposal We now propose our new linear-time dependence measure, the Finite Set Independence Criterion (FSIC). Let $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ be open sets. Let $\mu_x\mu_y(\mathbf{x},\mathbf{y}) := \mu_x(\mathbf{x})\mu_y(\mathbf{y})$. The idea is to see $\mu_{xy}(\mathbf{v},\mathbf{w}) = \mathbb{E}_{\mathbf{xy}}[k(\mathbf{x},\mathbf{v})l(\mathbf{y},\mathbf{w})]$, $\mu_x(\mathbf{v}) = \mathbb{E}_{\mathbf{x}}[k(\mathbf{x},\mathbf{v})]$ and $\mu_y(\mathbf{w}) = \mathbb{E}_{\mathbf{y}}[l(\mathbf{y},\mathbf{w})]$ as smooth functions, and consider a new distance between μ_{xy} and $\mu_x\mu_y$ instead of a Hilbert-Schmidt distance as in HSIC [Gretton et al., 2005a]. The new measure is given by the average of squared differences between μ_{xy} and $\mu_x\mu_y$, evaluated at J random test locations $V_J := \{(\mathbf{v}_i,\mathbf{w}_i)\}_{i=1}^J \subset \mathcal{X}\times\mathcal{Y}$.

$$\mathrm{FSIC}^2(X,Y) := \frac{1}{J}\sum_{i=1}^J u^2(\mathbf{v}_i,\mathbf{w}_i) = \frac{1}{J}\|\mathbf{u}\|_2^2,$$

where

$$\begin{aligned} u(\mathbf{v},\mathbf{w}) &:= \mu_{xy}(\mathbf{v},\mathbf{w}) - \mu_x(\mathbf{v})\mu_y(\mathbf{w}) \\ &= \mathbb{E}_{\mathbf{xy}}[k(\mathbf{x},\mathbf{v})l(\mathbf{y},\mathbf{w})] - \mathbb{E}_{\mathbf{x}}[k(\mathbf{x},\mathbf{v})]\mathbb{E}_{\mathbf{y}}[l(\mathbf{y},\mathbf{w})], \\ &= \mathrm{cov}_{\mathbf{xy}}[k(\mathbf{x},\mathbf{v}),l(\mathbf{y},\mathbf{w})], \end{aligned} \tag{4.2}$$

$\mathbf{u} := (u(\mathbf{v}_1,\mathbf{w}_1),\ldots,u(\mathbf{v}_J,\mathbf{w}_J))^\top$, and $\{(\mathbf{v}_i,\mathbf{w}_i)\}_{i=1}^J$ are realizations from an absolutely continuous distribution (wrt the Lebesgue measure).

Our first result in Proposition 4.2 states that $\mathrm{FSIC}(X,Y)$ almost surely defines a

dependence measure for the random variables X and Y, provided that the kernels k and l satisfy some conditions summarized in Assumption A.

Definition 4.1 (A_0 kernels)**.** Let $\mathcal{X}$ be an open set in $\mathbb{R}^d$. A positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be A_0 if it is bounded (i.e., there exists $B \in \mathbb{R}$ such that $\sup_{\mathbf{x},\mathbf{x}' \in \mathcal{X}} k(\mathbf{x}, \mathbf{x}') \leq B$), real analytic (Definition 2.11) and vanishes at infinity. Equivalently, for all $\mathbf{v} \in \mathcal{X}$, $f(\mathbf{x}) := k(\mathbf{x}, \mathbf{v})$ is bounded, real analytic on $\mathcal{X}$, and for all $\epsilon > 0$ the set $\{\mathbf{x} \mid |f(\mathbf{x})| \geq \epsilon\}$ is compact.[1]

Assumption A. The kernels $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $l\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ are A_0 (assumed to be bounded by B_k and B_l respectively), characteristic [Sriperumbudur et al., 2010, Definition 6], and translation invariant i.e., there exist $\check{k}$ and $\check{l}$ such that $k(\mathbf{x}, \mathbf{x}') = \check{k}(\mathbf{x} - \mathbf{x}')$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, and $l(\mathbf{y}, \mathbf{y}') = \check{l}(\mathbf{y} - \mathbf{y}')$ for all $\mathbf{y}, \mathbf{y}' \in \mathcal{Y}$.

Proposition 4.2 (FSIC is a dependence measure)**.** *Assume that assumption A holds, and that the test locations $V_J = \{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$ are drawn from an absolutely continuous distribution η. Then, η-almost surely,* $\mathrm{FSIC}(X, Y) = \frac{1}{\sqrt{J}}\|\mathbf{u}\|_2 = 0$ *if and only if X and Y are independent.*

Proof. To prove the backward direction, we note that if X and Y are independent, then $u(\mathbf{v}, \mathbf{w}) = 0$ for any $(\mathbf{v}, \mathbf{w})$ (see (4.2)), and $\mathrm{FSIC}(X, Y) = 0$. We will prove the forward direction. Define $u := \mu_{xy} - \mu_x \otimes \mu_y$, a member of the RKHS $\mathcal{H}_k \times \mathcal{H}_l$ associated with the product kernel $g((\mathbf{x}, \mathbf{y}), (\mathbf{v}, \mathbf{w})) := k(\mathbf{x}, \mathbf{v})l(\mathbf{y}, \mathbf{w})$. Since k and l are c_0-kernels (Definition 2.10), characteristic and translation invariant, Gretton [2015, Theorem 2] implies that $u = 0$ if and only if $P_{xy} = P_x P_y$. Since for all $\mathbf{v} \in \mathcal{X}, \mathbf{w} \in \mathcal{Y}$, $\mathbf{x} \mapsto k(\mathbf{x}, \mathbf{v})$ and $\mathbf{y} \mapsto l(\mathbf{y}, \mathbf{w})$ are real analytic, it follows from Lemma 4.9 that $(\mathbf{x}, \mathbf{y}) \mapsto k(\mathbf{x}, \mathbf{v})l(\mathbf{y}, \mathbf{w})$ is analytic on $\mathcal{X} \times \mathcal{Y}$. Since g is a bounded, real analytic kernel, Lemma 2.12 ensures that u is a real analytic function. It is known that if $u \neq 0$, the set of roots $R_u := \{(\mathbf{v}, \mathbf{w}) \mid u(\mathbf{v}, \mathbf{w}) = 0\}$ has Lebesgue measure zero [Mityagin, 2015]. Hence, it is sufficient to draw $(\mathbf{v}, \mathbf{w})$ from an absolutely continuous distribution to have $(\mathbf{v}, \mathbf{w}) \notin R_u$ η-almost surely, and consequently if X and Y are dependent, then $\mathrm{FSIC}(X, Y) > 0$, η-almost surely. □

Examples of kernels k and l which satisfy Assumption A are the Gaussian kernels. FSIC uses μ_{xy} as a proxy for P_{xy}, and $\mu_x \mu_y$ as a proxy for $P_x P_y$. Proposition 4.2 states that, to detect the dependence between X and Y, it is sufficient to evaluate the difference of the population joint embedding μ_{xy} and the embedding of the product of the marginal distributions $\mu_x \mu_y$ at a finite number of locations (defined by V_J). The intuitive explanation of this property is as follows. If $P_{xy} = P_x P_y$, then $u(\mathbf{v}, \mathbf{w}) = 0$ everywhere, and $\mathrm{FSIC}(X, Y) = 0$ for any V_J. If $P_{xy} \neq P_x P_y$, then u will not be a zero function. Using the same argument as in Chwialkowski et al. [2015], since k and l are analytic, u is also analytic, and the set of roots $R_u := \{(\mathbf{v}, \mathbf{w}) \mid u(\mathbf{v}, \mathbf{w}) = 0\}$

[1] A related class is the set of c_0 kernels [Sriperumbudur et al., 2010, Section 2]. A kernel k is said to be c_0 if it is bounded with $k(\cdot, \mathbf{v}) \in C_0(\mathcal{X})$ for all $\mathbf{v} \in \mathcal{X}$ where $C_0(\mathcal{X})$ contains the set of functions that vanish at infinity. Note that if k is A_0, then it is c_0. The converse is not necessarily true.

has Lebesgue measure zero [Mityagin, 2015]. Thus, it is sufficient to draw $(\mathbf{v}, \mathbf{w})$ from an absolutely continuous distribution η to have $(\mathbf{v}, \mathbf{w}) \notin R_u$ η-almost surely, and hence $\mathrm{FSIC}(X, Y) > 0$. We note that a characteristic kernel which is not analytic may produce u such that R_u has a positive Lebesgue measure. In this case, there is a positive probability that $(\mathbf{v}, \mathbf{w}) \in R_u$, resulting in a potential failure to detect the dependence. The required Assumption A only imposes conditions separately on each of the marginal kernels k and l. In particular, there is no condition on the product kernel $g((\mathbf{x}, \mathbf{y}), (\mathbf{v}, \mathbf{w})) := k(\mathbf{x}, \mathbf{v})l(\mathbf{y}, \mathbf{w})$ which would have been much more restrictive.

Plug-in Estimator Assume that we observe a joint sample $\mathsf{Z}_n := \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \overset{i.i.d.}{\sim} P_{xy}$. Unbiased estimators of $\mu_{xy}(\mathbf{v}, \mathbf{w})$ and $\mu_x\mu_y(\mathbf{v}, \mathbf{w})$ are $\hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) := \frac{1}{n}\sum_{i=1}^n k(\mathbf{x}_i, \mathbf{v})l(\mathbf{y}_i, \mathbf{w})$ and $\widehat{\mu_x\mu_y}(\mathbf{v}, \mathbf{w}) := \frac{1}{n(n-1)}\sum_{i=1}^n \sum_{j\neq i} k(\mathbf{x}_i, \mathbf{v})l(\mathbf{y}_j, \mathbf{w})$, respectively. A straightforward empirical estimator of FSIC^2 is then given by

$$\widehat{\mathrm{FSIC}^2}(\mathsf{Z}_n) = \frac{1}{J}\sum_{j=1}^J \hat{u}^2(\mathbf{v}_j, \mathbf{w}_j),$$

$$\hat{u}(\mathbf{v}, \mathbf{w}) := \hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) - \widehat{\mu_x\mu_y}(\mathbf{v}, \mathbf{w}) \tag{4.3}$$

$$= \frac{2}{n(n-1)}\sum_{i<j} h_{(\mathbf{v},\mathbf{w})}((\mathbf{x}_i, \mathbf{y}_i), (\mathbf{x}_j, \mathbf{y}_j)), \tag{4.4}$$

where

$$h_{(\mathbf{v},\mathbf{w})}((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')) := \frac{1}{2}(k(\mathbf{x}, \mathbf{v}) - k(\mathbf{x}', \mathbf{v}))(l(\mathbf{y}, \mathbf{w}) - l(\mathbf{y}', \mathbf{w})).$$

For conciseness, we define $\hat{\mathbf{u}} := (\hat{u}_1, \ldots, \hat{u}_J)^\top \in \mathbb{R}^J$ where $\hat{u}_i := \hat{u}(\mathbf{v}_i, \mathbf{w}_i)$ so that $\widehat{\mathrm{FSIC}^2}(\mathsf{Z}_n) = \frac{1}{J}\hat{\mathbf{u}}^\top\hat{\mathbf{u}}$.

$\widehat{\mathrm{FSIC}^2}$ can be efficiently computed in $\mathcal{O}((d_x + d_y)Jn)$ time which is linear in n [see (4.3) which does not have nested double sums], assuming that the runtime complexity of evaluating $k(\mathbf{x}, \mathbf{v})$ is $\mathcal{O}(d_x)$ and that of $l(\mathbf{y}, \mathbf{w})$ is $\mathcal{O}(d_y)$.

Since FSIC satisfies $\mathrm{FSIC}(X, Y) = 0 \iff X \perp Y$, in principle its empirical estimator can be used as a test statistic for an independence test proposing a null hypothesis H_0 : "X and Y are independent" against an alternative H_1 : "X and Y are dependent." The null distribution (i.e., distribution of the test statistic assuming that H_0 is true) is challenging to obtain, however, and depends on the unknown P_{xy}. This prompts us to consider a normalized version of FSIC whose asymptotic null distribution takes a more convenient form. We first derive the asymptotic distribution of $\hat{\mathbf{u}}$ in Proposition 4.3, which we use to derive the normalized test statistic in Theorem 4.4. As a shorthand, we write $\mathbf{z} := (\mathbf{x}, \mathbf{y})$, $\mathbf{t} := (\mathbf{v}, \mathbf{w})$, $\mathrm{cov}_{\mathbf{z}}$ is covariance,$\mathbb{V}_{\mathbf{z}}$ stands for variance.

Proposition 4.3 (Asymptotic distribution of $\hat{\mathbf{u}}$)**.** *Define* $\mathbf{u} := (u(\mathbf{t}_1), \ldots, u(\mathbf{t}_J))^\top$, $\tilde{k}(\mathbf{x}, \mathbf{v}) := k(\mathbf{x}, \mathbf{v}) - \mathbb{E}_{\mathbf{x}'}k(\mathbf{x}', \mathbf{v})$, *and* $\tilde{l}(\mathbf{y}, \mathbf{w}) := l(\mathbf{y}, \mathbf{w}) - \mathbb{E}_{\mathbf{y}'}l(\mathbf{y}', \mathbf{w})$. *Let* $\boldsymbol{\Sigma} = [\Sigma_{ij}] \in \mathbb{R}^{J\times J}$ *be the matrix such that*

$$\Sigma_{ij} = \mathbb{E}_{\mathbf{xy}}[\tilde{k}(\mathbf{x}, \mathbf{v}_i)\tilde{l}(\mathbf{y}, \mathbf{w}_i)\tilde{k}(\mathbf{x}, \mathbf{v}_j)\tilde{l}(\mathbf{y}, \mathbf{w}_j)] - u(\mathbf{t}_i)u(\mathbf{t}_j),$$

for $i, j \in \{1, \ldots, J\}$. Then, under both H_0 and H_1, for any fixed test locations $\{\mathbf{t}_1, \ldots, \mathbf{t}_J\}$ for which $\boldsymbol{\Sigma}$ is positive definite, it holds that $\sqrt{n}(\hat{\mathbf{u}} - \mathbf{u}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

Proof. For a fixed $\{\mathbf{t}_1, \ldots, \mathbf{t}_J\}$, $\hat{\mathbf{u}}$ is a one-sample second-order multivariate U-statistic with a U-statistic kernel $h_{\mathbf{t}}$ (see Section A.1: U-Statistics). Thus, by Lemma A.6, it follows directly that $\sqrt{n}(\hat{\mathbf{u}} - \mathbf{u}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ where we note that $\mathbb{E}_{\mathbf{xy}}[\tilde{k}(\mathbf{x}, \mathbf{v})\tilde{l}(\mathbf{y}, \mathbf{w})] = u(\mathbf{v}, \mathbf{w})$. □

Recall from Proposition 4.2 that $\mathbf{u} = \mathbf{0}$ holds almost surely under H_0. The asymptotic normality described in Proposition 4.3 implies that $n\widehat{\text{FSIC}^2} = \frac{n}{J}\hat{\mathbf{u}}^\top\hat{\mathbf{u}}$ converges in distribution to a sum of J dependent weighted χ^2 random variables. The dependence comes from the fact that the coordinates $\hat{u}_1 \ldots, \hat{u}_J$ of $\hat{\mathbf{u}}$ all depend on the sample Z_n. This null distribution requires a large number of simulations to compute the rejection threshold T_α for a given significance value α.

4.3 Normalized FSIC and Adaptive Test

For the purpose of an independence test, we will consider a normalized variant of $\widehat{\text{FSIC}^2}$, which we call $\widehat{\text{NFSIC}^2}$, whose tractable asymptotic null distribution is $\chi^2(J)$, the chi-squared distribution with J degrees of freedom. We then show that the independence test defined by $\widehat{\text{NFSIC}^2}$ is consistent. These results are given in Theorem 4.4.

Theorem 4.4 (Independence test based on $\widehat{\text{NFSIC}^2}$ is consistent)**.** *Let $\hat{\boldsymbol{\Sigma}}$ be a consistent estimate of $\boldsymbol{\Sigma}$ based on the joint sample Z_n, where $\boldsymbol{\Sigma}$ is defined in Proposition 4.3. Assume that $V_J = \{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J \sim \eta$ where η is absolutely continuous wrt the Lebesgue measure. The $\widehat{\text{NFSIC}^2}$ statistic is defined as $\hat{\lambda}_n := n\hat{\mathbf{u}}^\top \left(\hat{\boldsymbol{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}}$ where $\gamma_n \geq 0$ is a regularization parameter. Assume that*

1. *Assumption A holds.*
2. *$\boldsymbol{\Sigma}$ is invertible η-almost surely.*
3. *$\lim_{n\to\infty} \gamma_n = 0$.*

Then, for any k, l and V_J satisfying the assumptions above,

1. *Under H_0, $\hat{\lambda}_n \xrightarrow{d} \chi^2(J)$ as $n \to \infty$.*
2. *Under H_1, for any $r \in \mathbb{R}$, $\lim_{n\to\infty} \mathbb{P}\left(\hat{\lambda}_n \geq r\right) = 1$, η-almost surely.*

That is, the independence test based on $\widehat{\text{NFSIC}^2}$ is consistent.

Proof. Assume that H_0 holds. The consistency of $\hat{\boldsymbol{\Sigma}}$ and the continuous mapping theorem imply that $\left(\hat{\boldsymbol{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1} \xrightarrow{p} \boldsymbol{\Sigma}^{-1}$ which is a constant. Let $\mathbf{a}$ be a random vector in $\mathbb{R}^J$ following $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. By van der Vaart [2000, Theorem 2.7 (v)], it follows that $\left[\sqrt{n}\hat{\mathbf{u}}, \left(\hat{\boldsymbol{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1}\right] \xrightarrow{d} \left[\mathbf{a}, \boldsymbol{\Sigma}^{-1}\right]$ where $\mathbf{u} = 0$ almost surely by Proposition 4.2, and $\sqrt{n}\hat{\mathbf{u}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ by Proposition 4.3. Since $f(\mathbf{x}, \mathbf{S}) := \mathbf{x}^\top \mathbf{S}\mathbf{x}$ is continuous,

$f\left(\sqrt{n}\hat{\mathbf{u}}, \left(\hat{\boldsymbol{\Sigma}}+\gamma_n \mathbf{I}\right)^{-1}\right) \xrightarrow{d} f(\mathbf{a}, \boldsymbol{\Sigma}^{-1})$. Equivalently, $n\hat{\mathbf{u}}^\top \left(\hat{\boldsymbol{\Sigma}}+\gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}} \xrightarrow{d} \mathbf{a}^\top \boldsymbol{\Sigma}^{-1}\mathbf{a} \sim \chi^2(J)$ by Anderson [2003, Theorem 3.3.3]. This proves the first claim.

The proof of the second claim has a very similar structure to the proof of Proposition 2 of Chwialkowski et al. [2015]. Assume that H_1 holds. Then, $\mathbf{u} \neq \mathbf{0}$ almost surely by Proposition 4.2. Since k and l are bounded, it follows that $|h_{\mathbf{t}}(\mathbf{z}, \mathbf{z}')| \leq 2B_k B_l$ for any $\mathbf{z}, \mathbf{z}'$ (see (4.7)), and we have that $\hat{\mathbf{u}} \xrightarrow{a.s.} \mathbf{u}$ by Serfling [2009, Section 5.4, Theorem A]. Thus, $\hat{\mathbf{u}}^\top \left(\hat{\boldsymbol{\Sigma}}+\gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}} - \frac{r}{n} \xrightarrow{d} \mathbf{u}^\top \boldsymbol{\Sigma}^{-1}\mathbf{u}$ by the continuous mapping theorem, and the consistency of $\hat{\boldsymbol{\Sigma}}$. Consequently,

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}\left(\hat{\lambda}_n \geq r\right) &= 1 - \lim_{n\to\infty} \mathbb{P}\left(\hat{\mathbf{u}}^\top \left(\hat{\boldsymbol{\Sigma}}+\gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}} - \frac{r}{n} < 0\right) \\ &\overset{(a)}{=} 1 - \mathbb{P}\left(\mathbf{u}^\top \boldsymbol{\Sigma}^{-1}\mathbf{u} < 0\right) \overset{(b)}{=} 1,\end{aligned}$$

where at (a) we use the Portmanteau theorem [van der Vaart, 2000, Lemma 2.2 (i)] guaranteeing that $x_n \xrightarrow{d} x$ if and only if $\mathbb{P}(x_n < t) \to \mathbb{P}(x < t)$ for all continuity points of $t \mapsto \mathbb{P}(x < t)$. Step (b) is justified by noting that the covariance matrix $\boldsymbol{\Sigma}$ is positive definite so that $\mathbf{u}^\top \boldsymbol{\Sigma}^{-1}\mathbf{u} > 0$, and $t \mapsto \mathbb{P}(\mathbf{u}^\top \boldsymbol{\Sigma}^{-1}\mathbf{u} < t)$ (a step function) is continuous at 0. □

Theorem 4.4 states that if H_1 holds, the statistic can be arbitrarily large as n increases, allowing H_0 to be rejected for any threshold. Asymptotically the test threshold T_α is given by the $(1-\alpha)$-quantile of $\chi^2(J)$ and is independent of n. The assumption on the consistency of $\hat{\boldsymbol{\Sigma}}$ is required to obtain the asymptotic chi-squared distribution. The regularization parameter γ_n is to ensure that $(\hat{\boldsymbol{\Sigma}}+\gamma_n \mathbf{I})^{-1}$ can be stably computed. In practice, γ_n requires no tuning, and can be set to be a very small constant. We emphasize that J need not increase with n for test consistency.

The next proposition states that the computational complexity of the $\widehat{\text{NFSIC}^2}$ estimator is linear in both the input dimension and sample size, and that it can be expressed in terms of the $\mathbf{K} = [K_{ij}] = [k(\mathbf{v}_i, \mathbf{x}_j)] \in \mathbb{R}^{J\times n}, \mathbf{L} = [L_{ij}] = [l(\mathbf{w}_i, \mathbf{y}_j)] \in \mathbb{R}^{J\times n}$ matrices. In contrast to typical kernel methods, a large Gram matrix of size $n \times n$ is not needed to compute $\widehat{\text{NFSIC}^2}$.

Proposition 4.5 (An empirical estimator of $\widehat{\text{NFSIC}^2}$)**.** *Let* $\mathbf{1}_n := (1, \ldots, 1)^\top \in \mathbb{R}^n$. *Denote by* $\circ$ *the element-wise matrix product. Then,*

1. $\hat{\mathbf{u}} = \frac{(\mathbf{K}\circ\mathbf{L})\mathbf{1}_n}{n-1} - \frac{(\mathbf{K}\mathbf{1}_n)\circ(\mathbf{L}\mathbf{1}_n)}{n(n-1)}$.
2. *A consistent estimator for* $\boldsymbol{\Sigma}$ *is* $\hat{\boldsymbol{\Sigma}} = \frac{\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top}{n}$ *where*

$$\begin{aligned}\boldsymbol{\Gamma} &:= (\mathbf{K} - n^{-1}\mathbf{K}\mathbf{1}_n\mathbf{1}_n^\top) \circ (\mathbf{L} - n^{-1}\mathbf{L}\mathbf{1}_n\mathbf{1}_n^\top) - \hat{\mathbf{u}}^b\mathbf{1}_n^\top, \\ \hat{\mathbf{u}}^b &= n^{-1}(\mathbf{K}\circ\mathbf{L})\mathbf{1}_n - n^{-2}(\mathbf{K}\mathbf{1}_n)\circ(\mathbf{L}\mathbf{1}_n).\end{aligned}$$

Assume that the complexity of the kernel evaluation is linear in the input dimension. Then the test statistic $\hat{\lambda}_n = n\hat{\mathbf{u}}^\top \left(\hat{\boldsymbol{\Sigma}}+\gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}}$ *can be computed in* $\mathcal{O}(J^3 + J^2 n + (d_x + d_y)Jn)$ *time.*

Proof. Claim 1 for $\hat{\mathbf{u}}$ is straightforward. The expression for $\hat{\mathbf{\Sigma}}$ in claim 2 follows directly from the asymptotic covariance expression in Proposition 4.3. The consistency of $\hat{\mathbf{\Sigma}}$ can be obtained by noting that the finite sample bound for $\mathbb{P}(\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F > t)$ decreases as n increases. This is shown in Section 4.A. □

Although the dependency of the estimator on J is cubic, we empirically observe that only a small value of J is required (see Section 4.4). The number of test locations J relates to the number of regions in $\mathcal{X} \times \mathcal{Y}$ of p_{xy} and $p_x p_y$ that differ (see Figure 4.1).

Theorem 4.4 asserts the consistency of the test for any test locations V_J drawn from an absolutely continuous distribution. In practice, V_J can be further optimized to increase the test power for a fixed sample size. Our final theoretical result gives a lower bound on the test power of $\widehat{\text{NFSIC}^2}$ i.e., the probability of correctly rejecting H_0. We will use this lower bound as the objective function to determine V_J and the kernel parameters. Let $\|\cdot\|_F$ be the Frobenius norm.

Theorem 4.6 (A lower bound on the test power)**.** *Let* $\text{NFSIC}^2(X, Y) := \lambda_n := n\mathbf{u}^\top \mathbf{\Sigma}^{-1}\mathbf{u}$. *Let $\mathcal{K}$ be a kernel class for k, $\mathcal{L}$ be a kernel class for l, and $\mathcal{V}$ be a collection with each element being a set of J locations. Assume that*

1. *There exist finite B_k and B_l such that*

$$\sup_{k\in\mathcal{K}} \sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} |k(\mathbf{x},\mathbf{x}')| \leq B_k, \text{ and } \sup_{l\in\mathcal{L}} \sup_{\mathbf{y},\mathbf{y}'\in\mathcal{Y}} |l(\mathbf{y},\mathbf{y}')| \leq B_l.$$

2. $\tilde{c} := \sup_{k\in\mathcal{K}} \sup_{l\in\mathcal{L}} \sup_{V_J\in\mathcal{V}} \|\mathbf{\Sigma}^{-1}\|_F < \infty$.

Then, for any $k \in \mathcal{K}, l \in \mathcal{L}, V_J \in \mathcal{V}$, and $\lambda_n \geq r$, the test power satisfies $\mathbb{P}\left(\hat{\lambda}_n \geq r\right) \geq L(\lambda_n)$ where

$$L(\lambda_n) = 1 - 62e^{-\xi_1\gamma_n^2(\lambda_n - r)^2/n} - 2e^{-\lfloor 0.5n\rfloor(\lambda_n - r)^2/\left[\xi_2 n^2\right]} - 2e^{-\left[(\lambda_n - r)\gamma_n(n-1)/3 - \xi_3 n - c_3\gamma_n^2 n(n-1)\right]^2/\left[\xi_4 n^2(n-1)\right]},$$

$\lfloor\cdot\rfloor$ is the floor function, $\xi_1 := \frac{1}{3^2 c_1^2 J^2 B^*}$, B^* is a constant depending on only B_k and B_l, $\xi_2 := 72c_2^2 JB^2$, $B := B_k B_l$, $\xi_3 := 8c_1 B^2 J$, $c_3 := 4B^2 J\tilde{c}^2$, $\xi_4 := 2^8 B^4 J^2 c_1^2$, $c_1 := 4B^2 J\sqrt{J}\tilde{c}$, and $c_2 := 4B\sqrt{J}\tilde{c}$. Moreover, for sufficiently large fixed n, $L(\lambda_n)$ is increasing in λ_n.

We provide a proof in Section 4.A. To put Theorem 4.6 into perspective, assume that $\mathcal{K} = \left\{(\mathbf{x},\mathbf{v}) \mapsto \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma_x^2}\right) \mid \sigma_x^2 \in [\sigma_{x,l}^2, \sigma_{x,u}^2]\right\} =: \mathcal{K}_g$ for some $0 < \sigma_{x,l}^2 < \sigma_{x,u}^2 < \infty$ and $\mathcal{L} = \left\{(\mathbf{y},\mathbf{w}) \mapsto \exp\left(-\frac{\|\mathbf{y}-\mathbf{w}\|^2}{2\sigma_y^2}\right) \mid \sigma_y^2 \in [\sigma_{y,l}^2, \sigma_{y,u}^2]\right\} =: \mathcal{L}_g$ for some $0 < \sigma_{y,l}^2 < \sigma_{y,u}^2 < \infty$ are Gaussian kernel classes. Then, in Theorem 4.6, $B = B_k = B_l = 1$, and $B^* = 2$. The assumption $\tilde{c} < \infty$ is a technical condition to guarantee that the test power lower bound is finite for all θ defined by the feasible sets $\mathcal{K}, \mathcal{L}$, and $\mathcal{V}$. Let $\mathcal{V}_{\epsilon,r} := \left\{V_J \mid \|\mathbf{v}_i\|^2, \|\mathbf{w}_i\|^2 \leq r \text{ and } \|\mathbf{v}_i - \mathbf{v}_j\|_2^2 + \|\mathbf{w}_i - \mathbf{w}_j\|_2^2 \geq \epsilon, \text{ for all } i \neq j\right\}$. If we set $\mathcal{K} = \mathcal{K}_g, \mathcal{L} = \mathcal{L}_g$, and $\mathcal{V} = \mathcal{V}_{\epsilon,r}$ for some $\epsilon, r > 0$, then $\tilde{c} < \infty$ as $\mathcal{K}_g, \mathcal{L}_g$, and $\mathcal{V}_{\epsilon,r}$

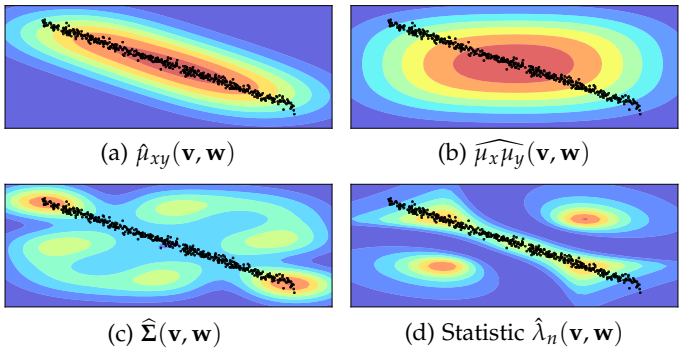
are compact. In practice, these conditions do not necessarily create restrictions as they almost always hold implicitly. We show in Section 4.4 that the objective function used to choose V_J will discourage any two locations to be in the same neighborhood.

Parameter Tuning Let θ be the collection of all tuning parameters of the test. If $k \in \mathcal{K}_g$ and $l \in \mathcal{L}_g$ (i.e., Gaussian kernels), then $\theta = \{\sigma_x^2, \sigma_y^2, V_J\}$. The test power lower bound $L(\lambda_n)$ in Theorem 4.6 is a function of $\lambda_n = n\mathbf{u}^\top \boldsymbol{\Sigma}^{-1}\mathbf{u}$ which is the population counterpart of the test statistic $\hat{\lambda}_n$. As in FSIC, it can be shown that $\lambda_n = 0$ if and only if X are Y are independent (from Proposition 4.2). According to Theorem 4.6, for a sufficiently large n, the test power lower bound is increasing in λ_n. One can therefore think of λ_n (a function of θ) as representing how easily the test rejects H_0 given a problem P_{xy}. The higher the λ_n, the greater the lower bound on the test power, and thus the more likely it is that the test will reject H_0 when it is false.

In light of this reasoning, we propose to set θ by maximizing the lower bound on the test power i.e., set θ to $\theta^* = \arg\max_\theta L(\lambda_n)$. Assume that n is sufficiently large so that $\lambda_n \mapsto L(\lambda_n)$ is an increasing function. Then, $\arg\max_\theta L(\lambda_n) = \arg\max_\theta \lambda_n$. Since λ_n is unknown, we propose dividing the sample Z_n into two disjoint sets: training and test sets. The training set is used to compute $\hat{\lambda}_n$ (an estimate of λ_n) to optimize for θ^*, and the test set is used for the actual independence test with the optimized θ^*. The splitting is to guarantee the independence of θ^* and the test sample to avoid overfitting. Since Theorem 4.4 guarantees that $\hat{\lambda}_n \xrightarrow{d} \chi^2(J)$ as $n \to \infty$ for any θ (that is independent of the test data), the asymptotic null distribution does not change by using θ^*. This implies that asymptotically the false rejection rate of H_0 (type-I error) is still at the design level of α.

Illustration of $\widehat{\text{NFSIC}}^2$ To better understand the behaviour of $\widehat{\text{NFSIC}}^2$, we visualize $\hat{\mu}_{xy}(\mathbf{v}, \mathbf{w})$, $\widehat{\mu_x\mu_y}(\mathbf{v}, \mathbf{w})$ and $\hat{\boldsymbol{\Sigma}}(\mathbf{v}, \mathbf{w})$ as a function of one test location $(\mathbf{v}, \mathbf{w})$ on a simple toy problem. In this problem, $Y = -X + Z$ where $Z \sim \mathcal{N}(0, 0.3^2)$ is an independent noise variable. As we consider only one location ($J = 1$), $\hat{\boldsymbol{\Sigma}}(\mathbf{v}, \mathbf{w})$ is a scalar. The statistic can be written as $\hat{\lambda}_n = n\frac{\left(\hat{\mu}_{xy}(\mathbf{v},\mathbf{w}) - \widehat{\mu_x\mu_y}(\mathbf{v},\mathbf{w})\right)^2}{\hat{\boldsymbol{\Sigma}}(\mathbf{v},\mathbf{w})}$. These components are shown in Figure 4.1, where we use Gaussian kernels for both X and Y, and the horizontal and vertical axes correspond to $\mathbf{v} \in \mathbb{R}$ and $\mathbf{w} \in \mathbb{R}$, respectively.

Intuitively, $\hat{u}(\mathbf{v}, \mathbf{w}) = \hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) - \widehat{\mu_x\mu_y}(\mathbf{v}, \mathbf{w})$ captures the difference of the joint distribution and the product of the marginals as a function of $(\mathbf{v}, \mathbf{w})$. Squaring $\hat{u}(\mathbf{v}, \mathbf{w})$ and dividing it by the variance shown in Figure 4.1c gives the statistic (also the tuning objective) shown in Figure 4.1d. The latter figure illustrates that the parameter tuning objective can be non-convex: non-convexity arises since there are multiple ways to detect the difference between the joint distribution and the product of the marginals. In this case, the lower left and upper right regions equally indicate the largest difference. A convex objective would not be able to capture this phenomenon.



(a) $\hat{\mu}_{xy}(\mathbf{v}, \mathbf{w})$ (b) $\widehat{\mu_x \mu_y}(\mathbf{v}, \mathbf{w})$

(c) $\widehat{\boldsymbol{\Sigma}}(\mathbf{v}, \mathbf{w})$ (d) Statistic $\hat{\lambda}_n(\mathbf{v}, \mathbf{w})$

Figure 4.1: Illustration of $\widehat{\text{NFSIC}}^2$. $Y = -X + Z$ where $Z \sim \mathcal{N}(0, 0.3^2)$ is an independent noise variable

4.4 Experiments

In this section, we empirically study the performance of the proposed method on both toy (Section 4.4.1) and real problems (Section 4.4.2). We are interested in challenging problems requiring a large number of samples, where a quadratic-time test might be computationally infeasible. Our goal is not to outperform a quadratic-time test with a linear-time test uniformly over *all* testing problems. We will find, however, that our test does outperform the quadratic-time test in some cases.

We compare the proposed NFSIC with optimization (NFSIC-opt) to five multivariate nonparametric tests. The $\widehat{\text{NFSIC}}^2$ test without optimization (NFSIC-med) acts as a baseline, allowing the effect of parameter optimization to be clearly seen. The original quadratic-time HSIC test of Gretton et al. [2005a] is denoted by QHSIC. Nyström HSIC (NyHSIC) uses a Nyström approximation to the kernel matrices of X and Y when computing the HSIC statistic. FHSIC is another variant of HSIC in which a random Fourier feature approximation [Rahimi and Recht, 2007] to the kernel is used. NyHSIC and FHSIC are studied in Zhang et al. [2017] and can be computed in $\mathcal{O}(n)$, with quadratic dependency on the number of inducing points in NyHSIC, and quadratic dependency on the number of random features in FHSIC. Finally, the Randomized Dependence Coefficient (RDC) proposed in Lopez-Paz et al. [2013] is also considered. The RDC can be seen as the primal form (with random Fourier features) of the kernel canonical correlation analysis of Bach and Jordan [2002] on copula-transformed data. We consider RDC as a linear-time test even though preprocessing by an empirical copula transform costs $\mathcal{O}((d_x + d_y)n \log n)$.

We use Gaussian kernel classes $\mathcal{K}_g$ and $\mathcal{L}_g$ for both X and Y in all the methods. Except NFSIC-opt, all other tests use full sample to conduct the independence test, where the Gaussian widths σ_x and σ_y are set according to the widely used median heuristic i.e., $\sigma_x = \text{median}\left(\left\{\|\mathbf{x}_i - \mathbf{x}_j\|_2 \mid 1 \leq i < j \leq n\right\}\right)$, and σ_y is set in the same way using $\{\mathbf{y}_i\}_{i=1}^n$. The J locations for NFSIC-med are randomly drawn from the standard multivariate normal distribution in each trial. For a sample of size n, NFSIC-opt uses half the sample for parameter tuning, and the other disjoint half for the

test. We permute the sample 300 times in RDC[2] and HSIC to simulate from the null distribution and compute the test threshold. The null distributions for FHSIC and NyHSIC are given by a finite sum of weighted $\chi^2(1)$ random variables given in Eq. 8 of Zhang et al. [2017]. Unless stated otherwise, we set the test threshold of the two NFSIC tests to be the $(1-\alpha)$-quantile of $\chi^2(J)$. To provide a fair comparison, we set $J = 10$, use 10 inducing points in NyHSIC, and 10 random Fourier features in FHSIC and RDC.

Optimization of NFSIC-opt The parameters of NFSIC-opt are σ_x, σ_y, and J locations of size $(d_x + d_y)J$. We treat all the parameters as a long vector in $\mathbb{R}^{2+(d_x+d_y)J}$ and use gradient ascent to optimize $\hat{\lambda}_{n/2}$. We observe that initializing V_J by randomly picking J points from the training sample yields good performance. The regularization parameter γ_n in NFSIC is fixed to a small value, and is not optimized. It is worth emphasizing that the complexity of the optimization procedure is still linear-time.[3] We do not consider the distance covariance (dCov) of Székely and Rizzo [2009] in the comparison since it was shown to be a special case of HSIC [Sejdinovic et al., 2013]. The asymptotically linear-time block HSIC test proposed in Zhang et al. [2017] is also omitted as FHSIC and NyHSIC have superior performance in their empirical study. Since FSIC, NyHFSIC and RDC rely on a finite-dimensional kernel approximation, these tests are consistent only if both the number of features increases with n. By constrast, the proposed NFSIC requires only n to go to infinity to achieve consistency i.e., J can be fixed.

4.4.1 Toy Problems

We consider three toy problems.

1. **Same Gaussian (SG).** The two variables are independently drawn from the standard multivariate normal distribution i.e., $X \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_x})$ and $Y \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_y})$ where $\mathbf{I}_d$ is the $d \times d$ identity matrix. This problem represents a case in which H_0 holds.

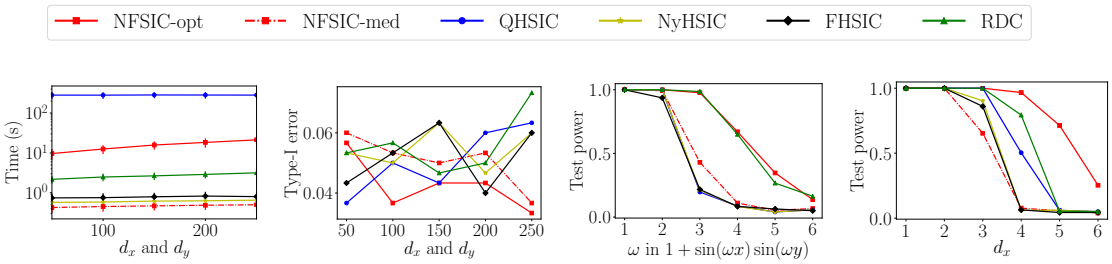
2. **Sinusoid (Sin).** Let p_{xy} be the probability density of P_{xy}. In the Sinusoid problem, the joint density of X and Y is given by

$$p_{xy}(x, y) \propto 1 + \sin(\omega x)\sin(\omega y), \tag{4.5}$$

   where the domains of $\mathcal{X} = (-\pi, \pi), \mathcal{Y} = (-\pi, \pi)$, and ω is the frequency of the sinusoid. As the frequency ω increases, the drawn sample becomes more similar to a sample drawn from Uniform$((-\pi, \pi)^2)$. That is, the higher ω, the harder to detect the dependency between X and Y. This problem was studied in Sejdinovic et al. [2013]. Plots of the density for a few values of ω are shown

[2] We use a permutation test for RDC, following the authors' implementation (https://github.com/lopezpaz/randomized_dependence_coefficient, referred commit: b0ac6c0).

[3] Our claim on linear runtime (with respect to n) is for the gradient ascent procedure to find a local optimum for θ. We do not claim a linear runtime to find a global optimum.



(a) SG ($\alpha = 0.05$) (b) SG ($\alpha = 0.05$) (c) Sin (d) GSign

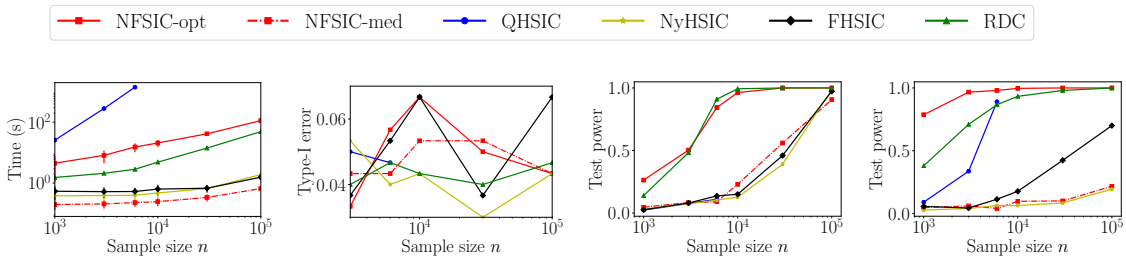
Figure 4.1: (a): Runtime. (b): Probability of rejecting H_0 as problem parameters vary. Fix $n = 4000$.

in Figures 4.4 and 4.5. The main characteristic of interest in this problem is the local change in the density function.

3. **Gaussian Sign (GSign).** In this problem, $Y = |Z| \prod_{i=1}^{d_x} \text{sgn}(X_i)$, where $X \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_x})$, $\text{sgn}(\cdot)$ is the sign function, and $Z \sim \mathcal{N}(0,1)$ serves as a source of noise. The full interaction of $X = (X_1, \ldots, X_{d_x})$ is what makes the problem challenging. That is, Y is dependent on X, yet it is independent of any proper subset of $\{X_1, \ldots, X_d\}$. Thus, simultaneous consideration of all the coordinates of X is required to successfully detect the dependency.

We fix $n = 4000$ and vary the problem parameters. Each problem is repeated for 300 trials, where a new sample is redrawn each time. The significance level α is set to 0.05. The results are shown in Figure 4.1. It can be seen that in the SG problem (Figure 4.1b) where H_0 holds, all the tests achieve roughly correct type-I errors at $\alpha = 0.05$. In particular, we point out that NFSIC-opt's rejection rate is well controlled as the sample used for testing and the sample used for parameter tuning are independent. The rejection rate would have been much higher had we done the optimization and testing on the same sample (i.e., overfitting). In the Sin problem, NFSIC-opt achieves high test power for all considered $\omega = 1, \ldots, 6$, highlighting its strength in detecting local changes in the joint density. The performance of NFSIC-med is significantly lower than that of NFSIC-opt. This phenomenon clearly emphasizes the importance of the optimization to place the locations at the relevant regions in $\mathcal{X} \times \mathcal{Y}$. RDC has a remarkably high performance in both Sin and GSign (Figure 4.1c, 4.1d) despite no parameter tuning. The ability to simultaneously consider interacting features of NFSIC-opt is indicated by its superior test power in GSign, especially at the challenging settings of $d_x = 5, 6$.

NFSIC vs. QHSIC. We observe that NFSIC-opt outperforms the quadratic-time QHSIC in these two problems. QHSIC is defined as the RKHS norm of the witness function u (see (4.2)). Intuitively, one can think of the RKHS norm as taking into account all the locations $(\mathbf{v}, \mathbf{w})$. By contrast, the proposed NFSIC evaluates the witness function at J locations. If the differences in p_{xy} and $p_x p_y$ are local (e.g., Sin problem), or there are interacting features (e.g., GSign problem), then only small regions in the space of (X, Y) are relevant in detecting the difference of p_{xy} and $p_x p_y$. In these



(a) SG. $d_x = d_y = 250$. (b) SG. $d_x = d_y = 250$. (c) Sin. $\omega = 4$. (d) GSign. $d_x = 4$.

Figure 4.2: (a) Runtime. (b): Probability of rejecting H_0 as n increases in the toy problems.

cases, pinpointing exact test locations by the optimization of NFSIC performs well. On the other hand, taking into account all possible test locations as done implicitly in QHSIC also integrates over regions where the difference between p_{xy} and $p_x p_y$ is small, resulting in a weaker indication of dependence. Whether QHSIC is better than NFSIC depends heavily on the problem, and there is no one best answer. If the difference between p_{xy} and $p_x p_y$ is large only in localized regions, then the proposed linear time statistic has an advantage. If the difference is spatially diffuse, then QHSIC has an advantage. No existing work has proposed a procedure to optimally tune kernel parameters for QHSIC; by contrast, NFSIC has a clearly defined objective for parameter tuning.

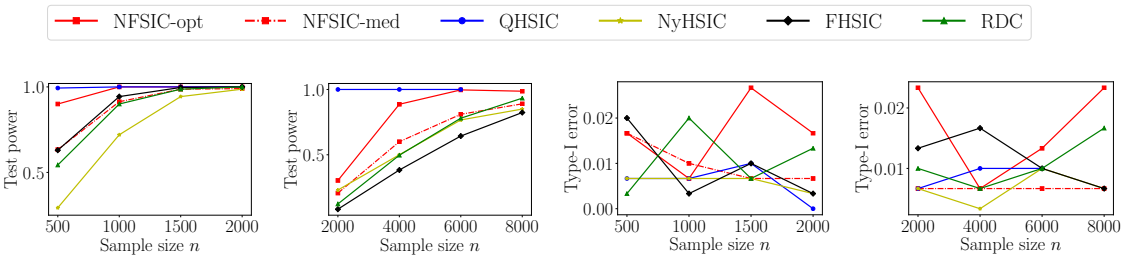
Sample Efficiency To investigate the sample efficiency of all the tests, we fix $d_x = d_y = 250$ in SG, $\omega = 4$ in Sin, $d_x = 4$ in GSign, and increase n. Figure 4.2 shows the results. The quadratic dependency on n in QHSIC makes it infeasible both in terms of memory and runtime to consider n larger than 6000 (Figure 4.2a). By contrast, although not the most time-efficient, NFSIC-opt has the highest sample-efficiency for GSign, and for Sin in the low-sample regime, significantly outperforming QHSIC. Despite the small additional overhead from the optimization, we are yet able to conduct an accurate test with $n = 10^5, d_x = d_y = 250$ in less than 100 seconds. We observe in Figure 4.2b that the two NFSIC variants have correct type-I errors across all sample sizes. We recall from Theorem 4.4 that the NFSIC test with random test locations will asymptotically reject H_0 if it is false. A demonstration of this property is given in Figure 4.2c, where the test power of NFSIC-med eventually reaches 1 with n higher than 10^5.

4.4.2 Real Problems

We now examine the performance of our proposed test on real problems.

Million Song Data (MSD) We consider a subset of the Million Song Data[4] [Bertin-Mahieux et al., 2011], in which each song (X) out of 515,345 is represented by 90 features, of which 12 features are timbre average (over all segments) of the song, and 78 features are timbre covariance. Most of the songs are western commercial tracks

[4]Million Song Data subset: https://archive.ics.uci.edu/ml/datasets/YearPredictionMSD.



(a) MSD (b) Videos & Captions (c) MSD (H_0) (d) Videos & Captions (H_0)

Figure 4.3: Probability of rejecting H_0 as n increases in the two real problems. $\alpha = 0.01$.

from 1922 to 2011. The goal is to detect the dependency between each song and its year of release (Y). We set $\alpha = 0.01$, and repeat for 300 trials where the full sample is randomly subsampled to n points in each trial. Other settings are the same as in the toy problems. To make sure that the type-I error is correct, we use the permutation approach in the NFSIC tests to compute the threshold. Figure 4.3a shows the test powers as n increases from 500 to 2000.

Evidently, NFSIC-opt has the highest test power among all the linear-time tests for all the sample sizes. Its test power is second to only QHSIC. We recall that NFSIC-opt uses half of the sample for parameter tuning. Thus, at $n = 500$, the actual sample for testing is 250, which is relatively small. The fact that there is a vast power gain from 0.4 (NFSIC-med) to 0.8 (NFSIC-opt) at $n = 500$ suggests that the optimization procedure can perform well even at a lower sample sizes.

Videos and Captions Our last problem is based on the VideoStory46K[5] dataset [Habibian et al., 2014]. The dataset contains 45,826 Youtube videos (X) of an average length of roughly one minute, and their corresponding text captions (Y) uploaded by the users. Each video is represented as a $d_x = 2000$ dimensional Fisher vector encoding of motion boundary histograms (MBH) descriptors of Wang and Schmid [2013]. Each caption is represented as a bag of words with each feature being the frequency of one word. After filtering only words which occur in at least six video captions, we obtain $d_y = 1878$ words. We examine the test powers as n increases from 2000 to 8000. The results are given in Figure 4.3. The problem is sufficiently challenging that all linear-time tests achieve a low power at $n = 2000$. QHSIC performs exceptionally well on this problem, achieving a maximum power throughout. NFSIC-opt has the highest sample efficiency among the linear-time tests, showing that the optimization procedure is also practical in a high dimensional setting.

Rejection Rate Under H_0 To simulate cases in which H_0 holds in the two real problems, we permute the sample to break the dependency of X and Y: for each $i \in \{1, \ldots, n\}$, pair $\mathbf{x}_i$ with $\mathbf{y}_j$ where j is randomly chosen such that $j \neq i$. The results are shown in Figure 4.3c and Figure 4.3d for the MSD and the Videos and Captions problems, respectively. We observe that all tests have correct false rejection rates (type-I errors) at the design level of $\alpha = 0.01$.

[5]VideoStory46K dataset: https://ivi.fnwi.uva.nl/isis/mediamill/datasets/videostory.php.

4.4.3 Redundant Test Locations

Here, we provide a simple illustration to show that two locations $\mathbf{t}_1 = (\mathbf{v}_1, \mathbf{w}_1)$ and $\mathbf{t}_2 = (\mathbf{v}_2, \mathbf{w}_2)$ which are too close to each other will reduce the maximization objective. We consider the Sinusoid problem described in (4.5) with $\omega = 1$ (see the joint density in the left figure of Figure (4.4)), and use $J = 2$ test locations. In Figure 4.4, $\mathbf{t}_1$ is fixed at the location indicated by the red marker, while $\mathbf{t}_2$ is varied along the horizontal line shown in green. The objective value $\hat{\lambda}_n$ as a function of $\mathbf{t}_2$ is shown in the bottom figure. It can be seen that $\hat{\lambda}_n$ decreases sharply when $\mathbf{t}_2$ is in the neighborhood of $\mathbf{t}_1$. This property implies that two locations which are too close will not maximize the objective function (i.e., the second feature contains no additional information when it matches the first). In general, for $J > 2$, the objective will sharply decrease if any two locations are in the same neighborhood.

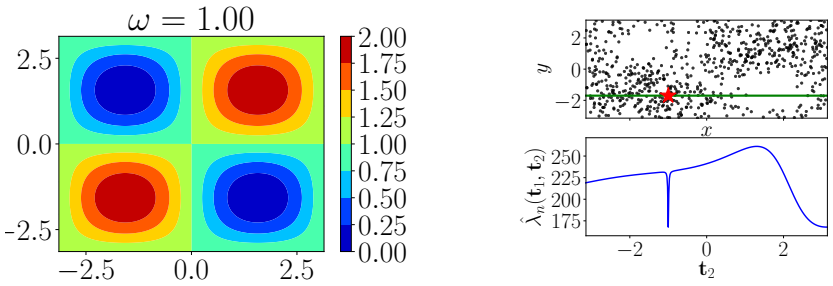


Figure 4.4: Plot of optimization objective values as location $\mathbf{t}_2$ moves along the green line. The objective sharply drops when the two locations are in the same neighborhood.

4.4.4 Test Power Vs. Number J of Test Locations

It might seem intuitive that as the number of locations J increases, the test power should also increase. Here, we empirically show that this statement is *not* always true. Consider the Sinusoid toy example described in (4.5) with $\omega = 2$ (also see the left figure of Figure 4.5). By construction, X and Y are dependent in this problem. We run NFSIC test with a sample size of $n = 800$, varying J from 1 to 600. For each value of J, the test is repeated for 500 times. In each trial, the sample is redrawn and the J test locations are drawn from Uniform$((-\pi, \pi)^2)$. There is no optimization of the test locations. We use Gaussian kernels for both X and Y, and use the median heuristic to set the Gaussian widths. Figure 4.5 shows the test power as J increases.

We observe that the test power does not monotonically increase as J increases. When $J = 1$, the difference of p_{xy} and $p_x p_y$ cannot be adequately captured, resulting in a low power. The power increases rapidly to roughly 0.6 at $J = 10$, and stays at 1 until about $J = 100$, after which the power steadily drops.

Unlike random Fourier features, the number of test locations in NFSIC is not the number of Monte Carlo particles used to approximate an expectation. There is a tradeoff: if the test locations are in key regions (i.e., regions in which there is a big difference between p_{xy} and $p_x p_y$), then they increase power; yet the statistic gains in

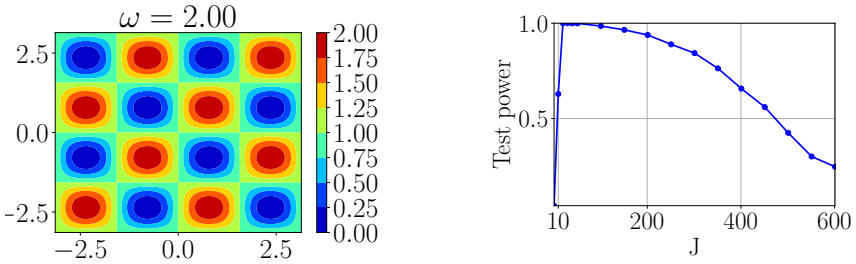


Figure 4.5: The Sinusoid problem and the plot of test power vs. the number of test locations.

variance (thus reducing test power) as J increases. As can be seen in Figure 4.5, there are 16 key regions (either bright red or bright blue) that can reveal the difference of p_{xy} and $p_x p_y$. Using an unnecessarily high J not only makes the covariance matrix $\hat{\mathbf{\Sigma}}$ harder to estimate accurately, it also significantly increases the computation as the complexity on J is $\mathcal{O}(J^3)$. We note that NFSIC is not intended to be used with a large J. In practice, it should be set to be large enough so as to capture the key regions as stated. As a practical guide, with optimization of the test locations, a good starting point is $J = 5$ or 10.

Proofs

4.A Proof: A Lower Bound on the Test Power

Recall Theorem 4.6,

Theorem 4.6 (A lower bound on the test power)**.** *Let* $\mathrm{NFSIC}^2(X,Y) := \lambda_n := n\mathbf{u}^\top \boldsymbol{\Sigma}^{-1}\mathbf{u}$. *Let* $\mathcal{K}$ *be a kernel class for* k, $\mathcal{L}$ *be a kernel class for* l, *and* $\mathcal{V}$ *be a collection with each element being a set of* J *locations. Assume that*

1. *There exist finite* B_k *and* B_l *such that*

$$\sup_{k\in\mathcal{K}}\sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}}|k(\mathbf{x},\mathbf{x}')| \leq B_k, \text{ and } \sup_{l\in\mathcal{L}}\sup_{\mathbf{y},\mathbf{y}'\in\mathcal{Y}}|l(\mathbf{y},\mathbf{y}')| \leq B_l.$$

2. $\tilde{c} := \sup_{k\in\mathcal{K}}\sup_{l\in\mathcal{L}}\sup_{V_J\in\mathcal{V}}\|\boldsymbol{\Sigma}^{-1}\|_F < \infty$.

Then, for any $k\in\mathcal{K}, l\in\mathcal{L}, V_J\in\mathcal{V}$, *and* $\lambda_n \geq r$, *the test power satisfies* $\mathbb{P}\left(\hat{\lambda}_n \geq r\right) \geq L(\lambda_n)$ *where*

$$\begin{aligned} L(\lambda_n) = 1 &- 62e^{-\xi_1\gamma_n^2(\lambda_n-r)^2/n} - 2e^{-\lfloor 0.5n\rfloor(\lambda_n-r)^2/\left[\xi_2 n^2\right]} \\ &- 2e^{-\left[(\lambda_n-r)\gamma_n(n-1)/3-\xi_3 n-c_3\gamma_n^2 n(n-1)\right]^2/\left[\xi_4 n^2(n-1)\right]}, \end{aligned}$$

$\lfloor\cdot\rfloor$ *is the floor function,* $\xi_1 := \frac{1}{3^2c_1^2J^2B^{*}}$, B^* *is a constant depending on only* B_k *and* B_l, $\xi_2 := 72c_2^2JB^2$, $B := B_kB_l$, $\xi_3 := 8c_1B^2J$, $c_3 := 4B^2J\tilde{c}^2$, $\xi_4 := 2^8B^4J^2c_1^2$, $c_1 := 4B^2J\sqrt{J}\tilde{c}$, *and* $c_2 := 4B\sqrt{J}\tilde{c}$. *Moreover, for sufficiently large fixed* n, $L(\lambda_n)$ *is increasing in* λ_n.

Overview of the proof We first derive a probabilistic bound for $|\hat{\lambda}_n - \lambda_n|/n$. The bound is in turn upper bounded by an expression involving $\|\hat{\mathbf{u}}-\mathbf{u}\|_2$ and $\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_F$. The difference $\|\hat{\mathbf{u}}-\mathbf{u}\|_2$ can be bounded by applying the bound for U-statistics given in Serfling [2009, Theorem A, p. 201]. For $\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_F$, we decompose it into a sum of smaller components, and bound each term with a product variant of the Hoeffding's inequality (Lemma 4.8). $L(\lambda_n)$ is obtained by combining all the bounds with the union bound.

Notations

Let $\langle\mathbf{A},\mathbf{B}\rangle_F := \mathrm{tr}(\mathbf{A}^\top\mathbf{B})$ denote the Frobenius inner product, and $\|\mathbf{A}\|_F := \sqrt{\mathrm{tr}(\mathbf{A}^\top\mathbf{A})}$ be the Frobenius norm. Write $\mathbf{z} := (\mathbf{x},\mathbf{y})$ to denote a pair of points from $\mathcal{X}\times\mathcal{Y}$.

We write $\mathbf{t} := (\mathbf{v}, \mathbf{w})$ to denote a pair of test locations from $\mathcal{X} \times \mathcal{Y}$. For brevity, an expectation over $(\mathbf{x}, \mathbf{y})$ (i.e., $\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P_{xy}}$) will be written as $\mathbb{E}_{\mathbf{z}}$ or $\mathbb{E}_{\mathbf{xy}}$. Define $\tilde{k}(\mathbf{x}, \mathbf{v}) := k(\mathbf{x}, \mathbf{v}) - \mathbb{E}_{\mathbf{x}'} k(\mathbf{x}', \mathbf{v})$, and $\tilde{l}(\mathbf{y}, \mathbf{w}) := l(\mathbf{y}, \mathbf{w}) - \mathbb{E}_{\mathbf{y}'} l(\mathbf{y}', \mathbf{w})$. Let $B_2(r) := \{\mathbf{x} \mid \|\mathbf{x}\|_2 \leq r\}$ be a closed ball with radius r centered at the origin. Similarly, define $B_F(r) := \{\mathbf{A} \mid \|\mathbf{A}\|_F \leq r\}$ to be a closed ball with radius r of $J \times J$ matrices under the Frobenius norm. Denote the max operation by $(x_1, \ldots, x_m)_+ = \max(x_1, \ldots, x_m)$.

For a product of marginal mean embeddings $\mu_x(\mathbf{v})\mu_y(\mathbf{w})$, we write $\widehat{\mu_x \mu_y}(\mathbf{v}, \mathbf{w}) := \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} k(\mathbf{x}_i, \mathbf{v}) l(\mathbf{y}_j, \mathbf{w})$ to denote the unbiased plug-in estimator, and write $\hat{\mu}_x(\mathbf{v})\hat{\mu}_y(\mathbf{w}) := \frac{1}{n} \sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v}) \frac{1}{n} \sum_{j=1}^{n} l(\mathbf{y}_j, \mathbf{w})$ which is a biased estimator. Define $\hat{u}^b(\mathbf{v}, \mathbf{w}) := \hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) - \hat{\mu}_x(\mathbf{v})\hat{\mu}_y(\mathbf{w})$ so that $\hat{\mathbf{u}}^b := \left(\hat{u}^b(\mathbf{t}_1), \ldots, \hat{u}^b(\mathbf{t}_J)\right)^\top$ where the superscript b stands for "biased". To avoid confusing with a positive definite kernel, we will refer to a U-statistic kernel as a *core*.

Proof

We will first derive a bound for $\mathbb{P}(|\hat{\lambda}_n - \lambda_n| \geq t)$, which will then be reparametrized to get a bound for the target quantity $\mathbb{P}(\hat{\lambda}_n \geq r)$. We closely follow the proof in Section 3.A.1 up to (4.11), then we diverge. We start by considering $|\hat{\lambda}_n - \lambda_n|/n$.

$$
\begin{aligned}
|\hat{\lambda}_n - \lambda_n|/n &= \left| \hat{\mathbf{u}}^\top (\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I})^{-1} \hat{\mathbf{u}} - \mathbf{u}^\top \mathbf{\Sigma}^{-1} \mathbf{u} \right| \\
&= \left| \hat{\mathbf{u}}^\top \left(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}} - \mathbf{u}^\top \left(\mathbf{\Sigma} + \gamma_n \mathbf{I}\right)^{-1} \mathbf{u} + \mathbf{u}^\top \left(\mathbf{\Sigma} + \gamma_n \mathbf{I}\right)^{-1} \mathbf{u} - \mathbf{u}^\top \mathbf{\Sigma}^{-1} \mathbf{u} \right| \\
&\leq \left| \hat{\mathbf{u}}^\top \left(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}} - \mathbf{u}^\top \left(\mathbf{\Sigma} + \gamma_n \mathbf{I}\right)^{-1} \mathbf{u} \right| + \left| \mathbf{u}^\top \left(\mathbf{\Sigma} + \gamma_n \mathbf{I}\right)^{-1} \mathbf{u} - \mathbf{u}^\top \mathbf{\Sigma}^{-1} \mathbf{u} \right| \\
&:= (\bigstar)_1 + (\bigstar)_2 .
\end{aligned}
$$

We next bound $(\bigstar_1)$ and $(\bigstar_2)$ separately by closely following the procedure in Section 3.A.1. We have

$$
\begin{aligned}
&\left| \hat{\mathbf{u}}^\top (\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I})^{-1} \hat{\mathbf{u}} - \mathbf{u}^\top \mathbf{\Sigma}^{-1} \mathbf{u} \right| \\
&\leq \frac{\sqrt{J}}{\gamma_n} \|\hat{\mathbf{u}}\|^2 \|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_F \|\mathbf{\Sigma}^{-1}\|_F + \left(\|\hat{\mathbf{u}}\|_2 + \|\mathbf{u}\|_2\right) \|\hat{\mathbf{u}} - \mathbf{u}\|_2 \|\mathbf{\Sigma}^{-1}\|_F + \gamma_n \|\mathbf{u}\|_2^2 \|\mathbf{\Sigma}^{-1}\|_F^2 .
\end{aligned} \tag{4.6}
$$

Bounding $\|\hat{\mathbf{u}}\|_2^2$ and $\|\mathbf{u}\|_2^2$ Here, we show that by the boundedness of the kernels k and l, it follows that $\|\hat{\mathbf{u}}\|_2^2$ is bounded. Recall that $\sup_{\mathbf{x},\mathbf{x}' \in \mathcal{X}} |k(\mathbf{x}, \mathbf{x}')| \leq B_k$, $\sup_{\mathbf{y},\mathbf{y}'} |l(\mathbf{y}, \mathbf{y}')| \leq B_l$, our notation $\mathbf{t} = (\mathbf{v}, \mathbf{w})$ for the test locations, and $\mathbf{z}_i := (\mathbf{x}_i, \mathbf{y}_i)$. We first show that the U-statistic core h is bounded.

$$
\begin{aligned}
\left| h_{\mathbf{t}}((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')) \right| &= \left| \frac{1}{2} (k(\mathbf{x}, \mathbf{v}) - k(\mathbf{x}', \mathbf{v}))(l(\mathbf{y}, \mathbf{w}) - l(\mathbf{y}', \mathbf{w})) \right| \\
&\leq \frac{1}{2} \left(|k(\mathbf{x}, \mathbf{v})| + |k(\mathbf{x}', \mathbf{v})| \right) \left(|l(\mathbf{y}, \mathbf{w})| + |l(\mathbf{y}', \mathbf{w})| \right) \\
&\leq 2 B_k B_l := 2B,
\end{aligned} \tag{4.7}
$$

where we define $B := B_k B_l$. It follows that

$$\|\hat{\mathbf{u}}\|_2^2 = \sum_{m=1}^{J} \left[\frac{2}{n(n-1)} \sum_{i<j} h_{\mathbf{t}_m}(\mathbf{z}_i, \mathbf{z}_j) \right]^2 \leq \sum_{m=1}^{J} [2B_k B_l]^2 = 4B^2 J, \tag{4.8}$$

$$\|\mathbf{u}\|_2^2 = \sum_{m=1}^{J} \left[\mathbb{E}_{\mathbf{z}} \mathbb{E}_{\mathbf{z}'} h_{\mathbf{t}_m}(\mathbf{z}, \mathbf{z}') \right]^2 \leq 4B^2 J. \tag{4.9}$$

Using the upper bounds on $\|\hat{\mathbf{u}}\|_2^2$, $\|\mathbf{u}\|_2^2$,(4.6) and the definition of $\tilde{c}$, we have

$$\begin{aligned} &\left| \hat{\mathbf{u}}^\top (\hat{\boldsymbol{\Sigma}} + \gamma_n \mathbf{I})^{-1} \hat{\mathbf{u}} - \mathbf{u}^\top \boldsymbol{\Sigma}^{-1} \mathbf{u} \right| \\ &\leq \frac{\sqrt{J}}{\gamma_n} 4B^2 J \tilde{c} \|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_F + 4B\sqrt{J}\tilde{c} \|\hat{\mathbf{u}} - \mathbf{u}\|_2 + 4B^2 J \tilde{c}^2 \gamma_n \\ &=: \frac{c_1}{\gamma_n} \|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_F + c_2 \|\hat{\mathbf{u}} - \mathbf{u}\|_2 + c_3 \gamma_n, \end{aligned} \tag{4.10}$$

where we define $c_1 := 4B^2 J\sqrt{J}\tilde{c}$, $c_2 := 4B\sqrt{J}\tilde{c}$, and $c_3 := 4B^2 J\tilde{c}^2$. This upper bound implies that

$$|\hat{\lambda}_n - \lambda_n| \leq \frac{c_1}{\gamma_n} n \|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_F + c_2 n \|\hat{\mathbf{u}} - \mathbf{u}\|_2 + c_3 n \gamma_n. \tag{4.11}$$

We will separately upper bound $\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_F$ and $\|\hat{\mathbf{u}} - \mathbf{u}\|_2$, and combine them with a union bound.

Bounding $\|\hat{\mathbf{u}} - \mathbf{u}\|_2$

Let $\mathbf{t}^* = \arg\max_{\mathbf{t} \in \{\mathbf{t}_1, \ldots, \mathbf{t}_J\}} |\hat{u}(\mathbf{t}) - u(\mathbf{t})|$. Recall that $\mathbf{u} = (u(\mathbf{t}_1), \ldots, u(\mathbf{t}_J))^\top = (u_1, \ldots, u_J)^\top$.

$$\begin{aligned} \|\hat{\mathbf{u}} - \mathbf{u}\|_2 &= \sup_{\mathbf{b} \in B_2(1)} \langle \mathbf{b}, \hat{\mathbf{u}} - \mathbf{u} \rangle_2 \leq \sup_{\mathbf{b} \in B_2(1)} \sum_{j=1}^{J} |b_j| |\hat{u}(\mathbf{t}_j) - u(\mathbf{t}_j)| \\ &\leq |\hat{u}(\mathbf{t}^*) - u(\mathbf{t}^*)| \sup_{\mathbf{b} \in B_2(1)} \sum_{j=1}^{J} |b_j| \\ &\overset{(a)}{\leq} \sqrt{J} |\hat{u}(\mathbf{t}^*) - u(\mathbf{t}^*)| \sup_{\mathbf{b} \in B_2(1)} \|\mathbf{b}\|_2 \\ &= \sqrt{J} |\hat{u}(\mathbf{t}^*) - u(\mathbf{t}^*)|, \end{aligned} \tag{4.12}$$

where at (a) we used $\|\mathbf{a}\|_1 \leq \sqrt{J}\|\mathbf{a}\|_2$ for any $\mathbf{a} \in \mathbb{R}^J$. From (4.12), it can be seen that bounding $\|\hat{\mathbf{u}} - \mathbf{u}\|_2$ amounts to bounding the difference of a U-statistic $\hat{u}(\mathbf{t}^*)$ (see (4.4)) to its expectation $u(\mathbf{t}^*)$. Combining (4.12) and (4.11), we have

$$|\hat{\lambda}_n - \lambda_n| \leq \frac{c_1}{\gamma_n} n \|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_F + c_2 n \sqrt{J} |\hat{u}(\mathbf{t}^*) - u(\mathbf{t}^*)| + c_3 n \gamma_n. \tag{4.13}$$

Bounding $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F$

The plan is to write $\hat{\mathbf{\Sigma}} = \hat{\mathbf{S}} - \hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top}, \mathbf{\Sigma} = \mathbf{S} - \mathbf{u}\mathbf{u}^\top$, so that $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F \leq \|\hat{\mathbf{S}} - \mathbf{S}\|_F + \|\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \mathbf{u}\mathbf{u}^\top\|_F$ and bound separately $\|\hat{\mathbf{S}} - \mathbf{S}\|_F$ and $\|\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \mathbf{u}\mathbf{u}^\top\|_F$.

Recall that $\Sigma_{ij} = \eta(\mathbf{t}_i, \mathbf{t}_j), \eta(\mathbf{t}, \mathbf{t}') = \mathbb{E}_{\mathbf{xy}}[\big(\tilde{k}(\mathbf{x}, \mathbf{v})\tilde{l}(\mathbf{y}, \mathbf{w}) - u(\mathbf{v}, \mathbf{w})\big)\big(\tilde{k}(\mathbf{x}, \mathbf{v}')\tilde{l}(\mathbf{y}, \mathbf{w}') - u(\mathbf{v}', \mathbf{w}')\big)]$ where $\tilde{k}(\mathbf{x}, \mathbf{v}) = k(\mathbf{x}, \mathbf{v}) - \mathbb{E}_{\mathbf{x}'}k(\mathbf{x}', \mathbf{v})$, and $\tilde{l}(\mathbf{y}, \mathbf{w}) = l(\mathbf{y}, \mathbf{w}) - \mathbb{E}_{\mathbf{y}'}l(\mathbf{y}', \mathbf{w})$. Its empirical estimator (see Proposition 4.5) is $\hat{\Sigma}_{ij} = \hat{\eta}(\mathbf{t}_i, \mathbf{t}_j)$ where

$$\begin{aligned}\hat{\eta}(\mathbf{t}, \mathbf{t}') &= \frac{1}{n}\sum_{i=1}^{n}[\big(\bar{k}(\mathbf{x}_i, \mathbf{v})\bar{l}(\mathbf{y}_i, \mathbf{w}) - \hat{u}^b(\mathbf{v}, \mathbf{w})\big)\big(\bar{k}(\mathbf{x}_i, \mathbf{v}')\bar{l}(\mathbf{y}_i, \mathbf{w}') - \hat{u}^b(\mathbf{v}', \mathbf{w}')\big)] \\ &= \frac{1}{n}\sum_{i=1}^{n}\bar{k}(\mathbf{x}_i, \mathbf{v})\bar{l}(\mathbf{y}_i, \mathbf{w})\bar{k}(\mathbf{x}_i, \mathbf{v}')\bar{l}(\mathbf{y}_i, \mathbf{w}') - \hat{u}^b(\mathbf{v}, \mathbf{w})\hat{u}^b(\mathbf{v}', \mathbf{w}'),\end{aligned}$$

$\bar{k}(\mathbf{x}, \mathbf{v}) := k(\mathbf{x}, \mathbf{v}) - \frac{1}{n}\sum_{i=1}^{n}k(\mathbf{x}_i, \mathbf{v})$, and $\bar{l}(\mathbf{y}, \mathbf{w}) := l(\mathbf{y}, \mathbf{w}) - \frac{1}{n}\sum_{i=1}^{n}l(\mathbf{y}_i, \mathbf{w})$. We note that $\frac{1}{n}\sum_{i=1}^{n}\bar{k}(\mathbf{x}_i, \mathbf{v})\bar{l}(\mathbf{y}_i, \mathbf{w}) = \hat{u}^b(\mathbf{v}, \mathbf{w})$. We define $\hat{\mathbf{S}} \in \mathbb{R}^{J\times J}$ such that

$$\hat{S}_{ij} := \frac{1}{n}\sum_{m=1}^{n}\bar{k}(\mathbf{x}_m, \mathbf{v}_i)\bar{l}(\mathbf{y}_m, \mathbf{w}_i)\bar{k}(\mathbf{x}_m, \mathbf{v}_j)\bar{l}(\mathbf{y}_i, \mathbf{w}_j),$$

and define similarly its population counterpart $\mathbf{S}$ such that

$$S_{ij} := \mathbb{E}_{\mathbf{xy}}[\tilde{k}(\mathbf{x}, \mathbf{v})\tilde{l}(\mathbf{y}, \mathbf{w})\tilde{k}(\mathbf{x}, \mathbf{v}')\tilde{l}(\mathbf{y}, \mathbf{w}')].$$

We have

$$\begin{aligned}\hat{\mathbf{\Sigma}} &= \hat{\mathbf{S}} - \hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top}, \\ \mathbf{\Sigma} &= \mathbf{S} - \mathbf{u}\mathbf{u}^\top, \\ \|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F &= \|\hat{\mathbf{S}} - \mathbf{S} - (\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \mathbf{u}\mathbf{u}^\top)\|_F \qquad (4.14) \\ &\leq \|\hat{\mathbf{S}} - \mathbf{S}\|_F + \|\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \mathbf{u}\mathbf{u}^\top\|_F. \qquad (4.15)\end{aligned}$$

With (4.15), (4.13) becomes

$$|\hat{\lambda}_n - \lambda_n| \leq \frac{c_1 n}{\gamma_n}\|\hat{\mathbf{S}} - \mathbf{S}\|_F + \frac{c_1 n}{\gamma_n}\|\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \mathbf{u}\mathbf{u}^\top\|_F + c_2 n\sqrt{J}|\hat{u}(\mathbf{t}^*) - u(\mathbf{t}^*)| + c_3 n\gamma_n. \qquad (4.16)$$

We will further separately bound $\|\hat{\mathbf{S}} - \mathbf{S}\|_F$ and $\|\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \mathbf{u}\mathbf{u}^\top\|_F$.

Bounding $\|\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \mathbf{u}\mathbf{u}^\top\|_F$

$$\begin{aligned}\|\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \mathbf{u}\mathbf{u}^\top\|_F &= \|\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \hat{\mathbf{u}}^b\mathbf{u}^\top + \hat{\mathbf{u}}^b\mathbf{u}^\top - \mathbf{u}\mathbf{u}^\top\|_F \\ &\leq \|\hat{\mathbf{u}}^b(\hat{\mathbf{u}}^b - \mathbf{u})^\top\|_F + \|(\hat{\mathbf{u}}^b - \mathbf{u})\mathbf{u}^\top\|_F \\ &= \|\hat{\mathbf{u}}^b\|_2\|\hat{\mathbf{u}}^b - \mathbf{u}\|_2 + \|\hat{\mathbf{u}}^b - \mathbf{u}\|_2\|\mathbf{u}\|_2 \\ &\leq 4B\sqrt{J}\|\hat{\mathbf{u}}^b - \mathbf{u}\|_2,\end{aligned}$$

where we used (4.9) and the fact that $\|\hat{\mathbf{u}}^b\|_2 \leq 2B\sqrt{J}$ which can be shown similarly to (4.8) as

$$\begin{aligned}\|\hat{\mathbf{u}}^b\|_2^2 &= \sum_{m=1}^{J} \left[\hat{\mu}_{xy}(\mathbf{v}_m, \mathbf{w}_m) - \hat{\mu}_x(\mathbf{v}_m)\hat{\mu}_y(\mathbf{w}_m)\right]^2 \\ &= \sum_{m=1}^{J} \left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} h_{\mathbf{t}_m}(\mathbf{z}_i, \mathbf{z}_j)\right]^2 \leq \sum_{m=1}^{J} [2B_k B_l]^2 = 4B^2 J.\end{aligned}$$

Let $(\tilde{\mathbf{v}}, \tilde{\mathbf{w}}) := \tilde{\mathbf{t}} = \arg\max_{\mathbf{t}\in\{\mathbf{t}_1,\ldots,\mathbf{t}_J\}} |\hat{u}^b(\mathbf{t}) - u(\mathbf{t})|$. We bound $\|\hat{\mathbf{u}}^b - \mathbf{u}\|_2$ by

$$\begin{aligned}\|\hat{\mathbf{u}}^b - \mathbf{u}\|_2 &\overset{(a)}{\leq} \sqrt{J}|\hat{u}^b(\tilde{\mathbf{t}}) - u(\tilde{\mathbf{t}})| \\ &= \sqrt{J}\left|\hat{\mu}_{xy}(\tilde{\mathbf{t}}) - \hat{\mu}_x(\tilde{\mathbf{v}})\hat{\mu}_y(\tilde{\mathbf{w}}) - u(\tilde{\mathbf{t}})\right| \\ &= \sqrt{J}\left|\hat{\mu}_{xy}(\tilde{\mathbf{t}}) - \widehat{\mu_x\mu_y}(\tilde{\mathbf{t}}) + \widehat{\mu_x\mu_y}(\tilde{\mathbf{t}}) - \hat{\mu}_x(\tilde{\mathbf{v}})\hat{\mu}_y(\tilde{\mathbf{w}}) - u(\tilde{\mathbf{t}})\right| \\ &\leq \sqrt{J}\left|\hat{\mu}_{xy}(\tilde{\mathbf{t}}) - \widehat{\mu_x\mu_y}(\tilde{\mathbf{t}}) - u(\tilde{\mathbf{t}})\right| + \sqrt{J}\left|\widehat{\mu_x\mu_y}(\tilde{\mathbf{t}}) - \hat{\mu}_x(\tilde{\mathbf{v}})\hat{\mu}_y(\tilde{\mathbf{w}})\right| \\ &= \sqrt{J}\,|\hat{u}(\tilde{\mathbf{t}}) - u(\tilde{\mathbf{t}})| + \sqrt{J}\left|\widehat{\mu_x\mu_y}(\tilde{\mathbf{t}}) - \hat{\mu}_x(\tilde{\mathbf{v}})\hat{\mu}_y(\tilde{\mathbf{w}})\right|, \qquad (4.17)\end{aligned}$$

where at (a) we used the same reasoning as in (4.12). The bias $\left|\widehat{\mu_x\mu_y}(\tilde{\mathbf{t}}) - \hat{\mu}_x(\tilde{\mathbf{v}})\hat{\mu}_y(\tilde{\mathbf{w}})\right|$ in the second term can be bounded as

$$\begin{aligned}&\left|\widehat{\mu_x\mu_y}(\tilde{\mathbf{t}}) - \hat{\mu}_x(\tilde{\mathbf{v}})\hat{\mu}_y(\tilde{\mathbf{w}})\right| \\ &= \left|\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} k(\mathbf{x}_i, \tilde{\mathbf{v}})l(\mathbf{y}_j, \tilde{\mathbf{w}}) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(\mathbf{x}_i, \tilde{\mathbf{v}})l(\mathbf{y}_j, \tilde{\mathbf{w}})\right| \\ &= \left|\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n} k(\mathbf{x}_i, \tilde{\mathbf{v}})l(\mathbf{y}_j, \tilde{\mathbf{w}}) - \frac{1}{n(n-1)}\sum_{i=1}^{n} k(\mathbf{x}_i, \tilde{\mathbf{v}})l(\mathbf{y}_i, \tilde{\mathbf{w}}) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(\mathbf{x}_i, \tilde{\mathbf{v}})l(\mathbf{y}_j, \tilde{\mathbf{w}})\right| \\ &= \left|\left(1 - \frac{n}{n-1}\right)\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(\mathbf{x}_i, \tilde{\mathbf{v}})l(\mathbf{y}_j, \tilde{\mathbf{w}}) + \frac{1}{n(n-1)}\sum_{i=1}^{n} k(\mathbf{x}_i, \tilde{\mathbf{v}})l(\mathbf{y}_i, \tilde{\mathbf{w}})\right| \\ &\leq \left|\left(1 - \frac{n}{n-1}\right)\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(\mathbf{x}_i, \tilde{\mathbf{v}})l(\mathbf{y}_j, \tilde{\mathbf{w}})\right| + \left|\frac{1}{n(n-1)}\sum_{i=1}^{n} k(\mathbf{x}_i, \tilde{\mathbf{v}})l(\mathbf{y}_i, \tilde{\mathbf{w}})\right| \\ &\leq \frac{B}{n-1} + \frac{B}{n-1} = \frac{2B}{n-1}.\end{aligned}$$

Combining this upper bound with (4.17), we have

$$\|\hat{\mathbf{u}}^b\hat{\mathbf{u}}^{b\top} - \mathbf{u}\mathbf{u}^\top\|_F \leq 4BJ\,|\hat{u}(\tilde{\mathbf{t}}) - u(\tilde{\mathbf{t}})| + \frac{8B^2J}{n-1}. \qquad (4.18)$$

With (4.18), (4.16) becomes

$$|\hat{\lambda}_n - \lambda_n| \leq \frac{c_1 n}{\gamma_n}\|\hat{\mathbf{S}} - \mathbf{S}\|_F + \frac{4BJc_1 n}{\gamma_n}|\hat{u}(\tilde{\mathbf{t}}) - u(\tilde{\mathbf{t}})| + \frac{c_1 n}{\gamma_n}\frac{8B^2J}{n-1} + c_2 n\sqrt{J}|\hat{u}(\mathbf{t}^*) - u(\mathbf{t}^*)| + c_3 n\gamma_n. \qquad (4.19)$$

Bounding $\|\hat{\mathbf{S}} - \mathbf{S}\|_F$

Recall that $V_J = \{\mathbf{t}_1, \ldots, \mathbf{t}_J\}$,

$$\hat{S}_{ij} = \hat{S}(\mathbf{t}_i, \mathbf{t}_j) = \frac{1}{n}\sum_{m=1}^{n} \bar{k}(\mathbf{x}_m, \mathbf{v}_i)\bar{l}(\mathbf{y}_m, \mathbf{w}_i)\bar{k}(\mathbf{x}_m, \mathbf{v}_j)\bar{l}(\mathbf{y}_m, \mathbf{w}_j),$$
$$S_{ij} = S(\mathbf{t}_i, \mathbf{t}_j) = \mathbb{E}_{\mathbf{xy}}[\tilde{k}(\mathbf{x}, \mathbf{v}_i)\tilde{l}(\mathbf{y}, \mathbf{w}_i)\tilde{k}(\mathbf{x}, \mathbf{v}_j)\tilde{l}(\mathbf{y}, \mathbf{w}_j)].$$

Let $(\mathbf{t}^{(1)}, \mathbf{t}^{(2)}) = \arg\max_{(\mathbf{s},\mathbf{t})\in V_J \times V_J} |\hat{S}(\mathbf{s}, \mathbf{t}) - S(\mathbf{s}, \mathbf{t})|$.

$$\begin{aligned}
\|\hat{\mathbf{S}} - \mathbf{S}\|_F &= \sup_{\mathbf{B}\in B_F(1)} \left\langle \mathbf{B}, \hat{\mathbf{S}} - \mathbf{S} \right\rangle_F \\
&\leq \sup_{\mathbf{B}\in B_F(1)} \sum_{i=1}^{J}\sum_{j=1}^{J} |B_{ij}||\hat{S}_{ij} - S_{ij}| \\
&\leq \left|\hat{S}(\mathbf{t}^{(1)}, \mathbf{t}^{(2)}) - S(\mathbf{t}^{(1)}, \mathbf{t}^{(2)})\right| \sup_{\mathbf{B}\in B_F(1)} \sum_{i=1}^{J}\sum_{j=1}^{J} |B_{ij}| \\
&\overset{(a)}{\leq} J\left|\hat{S}(\mathbf{t}^{(1)}, \mathbf{t}^{(2)}) - S(\mathbf{t}^{(1)}, \mathbf{t}^{(2)})\right| \sup_{\mathbf{B}\in B_F(1)} \|\mathbf{B}\|_F \\
&= J\left|\hat{S}(\mathbf{t}^{(1)}, \mathbf{t}^{(2)}) - S(\mathbf{t}^{(1)}, \mathbf{t}^{(2)})\right|, \qquad (4.20)
\end{aligned}$$

where at (a) we used $\sum_{i=1}^{J}\sum_{j=1}^{J} |A_{ij}| \leq J\|\mathbf{A}\|_F$ for any matrix $\mathbf{A} \in \mathbb{R}^{J\times J}$. We arrive at

$$\begin{aligned}
|\hat{\lambda}_n - \lambda_n| \leq \frac{c_1 J n}{\gamma_n}\left|\hat{S}(\mathbf{t}^{(1)}, \mathbf{t}^{(2)}) - S(\mathbf{t}^{(1)}, \mathbf{t}^{(2)})\right| + \frac{4BJc_1 n}{\gamma_n}|\hat{u}(\tilde{\mathbf{t}}) - u(\tilde{\mathbf{t}})| \\
+ \frac{c_1 n}{\gamma_n}\frac{8B^2 J}{n-1} + c_2 n\sqrt{J}|\hat{u}(\mathbf{t}^*) - u(\mathbf{t}^*)| + c_3 n\gamma_n. \qquad (4.21)
\end{aligned}$$

Bounding $\left|\hat{S}(\mathbf{t}, \mathbf{t}') - S(\mathbf{t}, \mathbf{t}')\right|$

Having an upper bound for $\left|\hat{S}(\mathbf{t}, \mathbf{t}') - S(\mathbf{t}, \mathbf{t}')\right|$ will allow us to bound (4.21). To keep the notations uncluttered, we will define the following shorthands.

Expression	Shorthand
$k(\mathbf{x}, \mathbf{v})$	a
$k(\mathbf{x}, \mathbf{v}')$	a'
$k(\mathbf{x}_i, \mathbf{v})$	a_i
$k(\mathbf{x}_i, \mathbf{v}')$	a_i'
$\mathbb{E}_{\mathbf{x}\sim P_x} k(\mathbf{x}, \mathbf{v})$	$\tilde{a}$
$\mathbb{E}_{\mathbf{x}\sim P_x} k(\mathbf{x}, \mathbf{v}')$	$\tilde{a}'$
$\frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v})$	$\bar{a}$
$\frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v}')$	$\bar{a}'$

Expression	Shorthand
$l(\mathbf{y}, \mathbf{w})$	b
$l(\mathbf{y}, \mathbf{w}')$	b'
$l(\mathbf{y}_i, \mathbf{w})$	b_i
$l(\mathbf{y}_i, \mathbf{w}')$	b_i'
$\mathbb{E}_{\mathbf{y}\sim P_y} l(\mathbf{y}, \mathbf{w})$	$\tilde{b}$
$\mathbb{E}_{\mathbf{y}\sim P_y} l(\mathbf{y}, \mathbf{w}')$	$\tilde{b}'$
$\frac{1}{n}\sum_{i=1}^{n} l(\mathbf{y}_i, \mathbf{w})$	$\bar{b}$
$\frac{1}{n}\sum_{i=1}^{n} l(\mathbf{y}_i, \mathbf{w}')$	$\bar{b}'$

We will also use $\overline{\cdot}$ to denote a empirical expectation over $\mathbf{x}$, or $\mathbf{y}$, or $(\mathbf{x}, \mathbf{y})$. The argument under $\overline{\cdot}$ will determine the variable over which we take the expectation. For instance, $\overline{aa'} = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v})k(\mathbf{x}_i, \mathbf{v}')$ and $\overline{aba'} = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v})l(\mathbf{y}_i, \mathbf{w})k(\mathbf{x}_i, \mathbf{v}')$, and so on. We define in the same way for the population expectation using $\widetilde{\cdot}$ i.e., $\widetilde{aa'} = \mathbb{E}_{\mathbf{x}}\left[k(\mathbf{x}, \mathbf{v})k(\mathbf{x}, \mathbf{v}')\right]$ and $\widetilde{aba'} = \mathbb{E}_{\mathbf{xy}}\left[k(\mathbf{x}, \mathbf{v})l(\mathbf{y}, \mathbf{w})k(\mathbf{x}, \mathbf{v}')\right]$.

With these shorthands, we can rewrite $\hat{S}(\mathbf{t}, \mathbf{t}')$ and $S(\mathbf{t}, \mathbf{t}')$ as

$$\hat{S}(\mathbf{t}, \mathbf{t}') = \frac{1}{n}\sum_{i=1}^{n}(a_i - \overline{a})(b_i - \overline{b})(a_i' - \overline{a}')(b_i' - \overline{b}'),$$
$$S(\mathbf{t}, \mathbf{t}') = \mathbb{E}_{\mathbf{xy}}\left[(a - \tilde{a})(b - \tilde{b})(a' - \tilde{a}')(b' - \tilde{b}')\right].$$

By expanding $S(\mathbf{t}, \mathbf{t}')$, we have

$$\begin{aligned}
S(\mathbf{t}, \mathbf{t}') = \mathbb{E}_{\mathbf{xy}}\big[&+ aba'b' - aba'\tilde{b}' - ab\tilde{a}'b' + ab\tilde{a}'\tilde{b}' \\
&- a\tilde{b}a'b' + a\tilde{b}a'\tilde{b}' + a\tilde{b}\tilde{a}'b' - a\tilde{b}\tilde{a}'\tilde{b}' \\
&- \tilde{a}ba'b' + \tilde{a}ba'\tilde{b}' + \tilde{a}b\tilde{a}'b' - \tilde{a}b\tilde{a}'\tilde{b}' \\
&+ \tilde{a}\tilde{b}a'b' - \tilde{a}\tilde{b}a'\tilde{b}' - \tilde{a}\tilde{b}\tilde{a}'\tilde{b}' + \tilde{a}\tilde{b}\tilde{a}'\tilde{b}'\big] \\
= &+\widetilde{aba'b'} - \widetilde{aba'\tilde{b}'} - \widetilde{abb'\tilde{a}'} + \widetilde{ab\tilde{a}'\tilde{b}'} \\
&- \widetilde{aa'b'\tilde{b}} + \widetilde{aa'\tilde{b}\tilde{b}'} + \widetilde{ab'\tilde{a}'\tilde{b}} - \tilde{a}\tilde{b}\tilde{a}'\tilde{b}' \\
&- \widetilde{a'bb'\tilde{a}} + \widetilde{a'b\tilde{a}\tilde{b}'} + \widetilde{\tilde{a}\tilde{a}'bb'} - \tilde{a}\tilde{b}\tilde{a}'\tilde{b}' \\
&+ \widetilde{a'b'\tilde{a}\tilde{b}} - \tilde{a}\tilde{b}\tilde{a}'\tilde{b}' - \tilde{a}\tilde{b}\tilde{a}'\tilde{b}' + \tilde{a}\tilde{b}\tilde{a}'\tilde{b}' \\
= &+\widetilde{aba'b'} - \widetilde{aba'\tilde{b}'} - \widetilde{abb'\tilde{a}'} + \widetilde{ab\tilde{a}'\tilde{b}'} \\
&- \widetilde{aa'b'\tilde{b}} + \widetilde{aa'\tilde{b}\tilde{b}'} + \widetilde{ab'\tilde{a}'\tilde{b}} + \widetilde{a'b'\tilde{a}\tilde{b}} \\
&- \widetilde{a'bb'\tilde{a}} + \widetilde{a'b\tilde{a}\tilde{b}'} + \widetilde{\tilde{a}\tilde{a}'bb'} - 3\tilde{a}\tilde{b}\tilde{a}'\tilde{b}'.
\end{aligned}$$

The expansion of $\hat{S}(\mathbf{t}, \mathbf{t}')$ can be done in the same way. By the triangle inequality, we have

$$\begin{aligned}
&\left|\hat{S}(\mathbf{t}, \mathbf{t}') - S(\mathbf{t}, \mathbf{t}')\right| \\
&\leq \left|\overline{aba'b'} - \widetilde{aba'b'}\right| + \left|\overline{aba'}\,\overline{b}' - \widetilde{aba'\tilde{b}'}\right| + \left|\overline{abb'}\overline{a}' - \widetilde{abb'\tilde{a}'}\right| + \left|\overline{ab\tilde{a}'}\overline{b}' - \widetilde{ab\tilde{a}'\tilde{b}'}\right| \\
&\quad \left|\overline{aa'b'}\,\overline{b} - \widetilde{aa'b'\tilde{b}}\right| + \left|\overline{aa'}\,\overline{b}\,\overline{b}' - \widetilde{aa'\tilde{b}\tilde{b}'}\right| + \left|\overline{ab'}\overline{a}'\overline{b} - \widetilde{ab'\tilde{a}'\tilde{b}}\right| + \left|\overline{a'b'}\overline{a}\overline{b} - \widetilde{a'b'\tilde{a}\tilde{b}}\right| \\
&\quad \left|\overline{a'bb'}\overline{a} - \widetilde{a'bb'\tilde{a}}\right| + \left|\overline{a'b\tilde{a}}\overline{b}' - \widetilde{a'b\tilde{a}\tilde{b}'}\right| + \left|\overline{a}\,\overline{a}'\overline{bb'} - \widetilde{\tilde{a}\tilde{a}'bb'}\right| + 3\left|\overline{a\tilde{b}\tilde{a}'}\overline{b}' - \widetilde{\tilde{a}\tilde{b}\tilde{a}'\tilde{b}'}\right|.
\end{aligned}$$

The first term $\left|\overline{aba'b'} - \widetilde{aba'b'}\right|$ can be bounded by applying the Hoeffding's inequality. Other terms can be bounded by applying Lemma 4.8. Recall that we write $(x_1, \ldots, x_m)_+$ for $\max(x_1, \ldots, x_m)$.

Bounding $\left|\overline{aba'b'} - \widetilde{aba'b'}\right|$ (1^{st} term). Since $-B^2 \leq aba'b' \leq B^2$, by the Hoeffding's inequality (Lemma 3.13), we have

$$\mathbb{P}\left(\left|\overline{aba'b'} - \widetilde{aba'b'}\right| \leq t\right) \geq 1 - 2\exp\left(-\frac{nt^2}{2B^4}\right).$$

Bounding $\left|\overline{aba'}\,\overline{b}' - \widetilde{aba'}\tilde{b}'\right|$ (2^{nd} term). Let $f_1(\mathbf{x},\mathbf{y}) = aba' = k(\mathbf{x},\mathbf{v})l(\mathbf{y},\mathbf{w})k(\mathbf{x},\mathbf{v}')$ and $f_2(\mathbf{y}) = b' = l(\mathbf{y},\mathbf{w}')$. We note that $|f_1(\mathbf{x},\mathbf{y})| \leq (BB_k, B_l)_+$ and $|f_2(\mathbf{y})| \leq (BB_k, B_l)_+$. Thus, by Lemma 4.8 with $E = 2$, we have

$$\mathbb{P}\left(\left|\overline{aba'}\,\overline{b}' - \widetilde{aba'}\tilde{b}'\right| \leq t\right) \geq 1 - 4\exp\left(-\frac{nt^2}{8(BB_k, B_l)_+^4}\right).$$

Bounding $\left|\overline{ab}\overline{a}'\overline{b}' - \widetilde{ab}\tilde{a}'\tilde{b}'\right|$ (4^{th} term). Let $f_1(\mathbf{x},\mathbf{y}) = ab = k(\mathbf{x},\mathbf{v})l(\mathbf{y},\mathbf{w})$, $f_2(\mathbf{x}) = a' = k(\mathbf{x},\mathbf{v}')$ and $f_3(\mathbf{y}) = b' = l(\mathbf{y},\mathbf{w}')$. We can see that $|f_1(\mathbf{x},\mathbf{y})|, |f_2(\mathbf{x})|, |f_3(\mathbf{y})| \leq (B, B_k, B_l)_+$. Thus, by Lemma 4.8 with $E = 3$, we have

$$\mathbb{P}\left(\left|\overline{ab}\overline{a}'\overline{b}' - \widetilde{ab}\tilde{a}'\tilde{b}'\right| \leq t\right) \geq 1 - 6\exp\left(-\frac{nt^2}{18(B, B_k, B_l)_+^6}\right).$$

Bounding $\left|\bar{a}\bar{b}\bar{a}'\bar{b}' - \tilde{a}\tilde{b}\tilde{a}'\tilde{b}'\right|$ (last term). Let $f_1(\mathbf{x}) = a = k(\mathbf{x},\mathbf{v}), f_2(\mathbf{y}) = b = l(\mathbf{y},\mathbf{w}), f_3(\mathbf{x}) = a' = k(\mathbf{x},\mathbf{v}')$ and $f_4(\mathbf{y}) = b' = l(\mathbf{y},\mathbf{w}')$. It can be seen that $|f_1(\mathbf{x})|, |f_2(\mathbf{y})|, |f_3(\mathbf{x})|, |f_4(\mathbf{y})| \leq (B_k, B_l)_+$. Thus, by Lemma 4.8 with $E = 4$, we have

$$\mathbb{P}\left(3\left|\bar{a}\bar{b}\bar{a}'\bar{b}' - \tilde{a}\tilde{b}\tilde{a}'\tilde{b}'\right| \leq t\right) \geq 1 - 8\exp\left(-\frac{nt^2}{32\cdot 3^2(B_k, B_l)_+^8}\right).$$

Bounds for other terms can be derived in a similar way to yield

$$(3^{rd}\text{ term})\quad \mathbb{P}\left(\left|\overline{abb'}\overline{a}' - \widetilde{abb'}\tilde{a}'\right| \leq t\right) \geq 1 - 4\exp\left(-\frac{nt^2}{8(BB_l, B_k)_+^4}\right),$$

$$(5^{th}\text{ term})\quad \mathbb{P}\left(\left|\overline{aa'b'}\,\overline{b} - \widetilde{aa'b'}\tilde{b}\right| \leq t\right) \geq 1 - 4\exp\left(-\frac{nt^2}{8(BB_k, B_l)_+^4}\right),$$

$$(6^{th}\text{ term})\quad \mathbb{P}\left(\left|\overline{aa'}\,\overline{b}\,\overline{b}' - \widetilde{aa'}\tilde{b}\tilde{b}'\right| \leq t\right) \geq 1 - 6\exp\left(-\frac{nt^2}{18(B_k^2, B_l)_+^6}\right),$$

$$(7^{th}\text{ term})\quad \mathbb{P}\left(\left|\overline{ab'}\overline{a}'\overline{b} - \widetilde{ab'}\tilde{a}'\tilde{b}\right| \leq t\right) \geq 1 - 6\exp\left(-\frac{nt^2}{18(B, B_k, B_l)_+^6}\right),$$

$$(8^{th}\text{ term})\quad \mathbb{P}\left(\left|\overline{a'b'}\overline{a}\overline{b} - \widetilde{a'b'}\tilde{a}\tilde{b}\right| \leq t\right) \geq 1 - 6\exp\left(-\frac{nt^2}{18(B, B_k, B_l)_+^6}\right),$$

$$(9^{th}\text{ term})\quad \mathbb{P}\left(\left|\overline{a'bb'}\overline{a} - \widetilde{a'bb'}\tilde{a}\right| \leq t\right) \geq 1 - 4\exp\left(-\frac{nt^2}{8(BB_l, B_k)_+^4}\right),$$

$$(10^{th}\text{ term})\quad \mathbb{P}\left(\left|\overline{a'b}\overline{a}\overline{b}' - \widetilde{a'b}\tilde{a}\tilde{b}'\right| \leq t\right) \geq 1 - 6\exp\left(-\frac{nt^2}{18(B, B_k, B_l)_+^6}\right),$$

$$(11^{th}\text{ term})\quad \mathbb{P}\left(\left|\overline{a}\,\overline{a}'\overline{bb'}-\tilde{a}\tilde{a}'\widetilde{bb'}\right|\le t\right)\ge 1-6\exp\left(-\frac{nt^2}{18(B_k,B_l^2)_+^6}\right).$$

By the union bound, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{S}(\mathbf{t},\mathbf{t}')-S(\mathbf{t},\mathbf{t}')\right|\le 12t\right)\\
&\ge 1-\left[2\exp\left(-\frac{nt^2}{2B^4}\right)+4\exp\left(-\frac{nt^2}{8(BB_k,B_l)_+^4}\right)+4\exp\left(-\frac{nt^2}{8(BB_l,B_k)_+^4}\right)+6\exp\left(-\frac{nt^2}{18(B,B_k,B_l)_+^6}\right)\right.\\
&\quad+4\exp\left(-\frac{nt^2}{8(BB_k,B_l)_+^4}\right)+6\exp\left(-\frac{nt^2}{18(B_k^2,B_l)_+^6}\right)+6\exp\left(-\frac{nt^2}{18(B,B_k,B_l)_+^6}\right)\\
&\quad+6\exp\left(-\frac{nt^2}{18(B,B_k,B_l)_+^6}\right)+4\exp\left(-\frac{nt^2}{8(BB_l,B_k)_+^4}\right)+6\exp\left(-\frac{nt^2}{18(B,B_k,B_l)_+^6}\right)\\
&\quad\left.+6\exp\left(-\frac{nt^2}{18(B_k,B_l^2)_+^6}\right)+8\exp\left(-\frac{nt^2}{32\cdot 3^2(B_k,B_l)_+^8}\right)\right]\\
&=1-\left[2\exp\left(-\frac{nt^2}{2B^4}\right)+8\exp\left(-\frac{nt^2}{8(BB_k,B_l)_+^4}\right)+8\exp\left(-\frac{nt^2}{8(BB_l,B_k)_+^4}\right)+24\exp\left(-\frac{nt^2}{18(B,B_k,B_l)_+^6}\right)\right.\\
&\quad\left.+6\exp\left(-\frac{nt^2}{18(B_k^2,B_l)_+^6}\right)+6\exp\left(-\frac{nt^2}{18(B_k,B_l^2)_+^6}\right)+8\exp\left(-\frac{nt^2}{32\cdot 3^2(B_k,B_l)_+^8}\right)\right]\\
&\ge 1-\left[2\exp\left(-\frac{12^2nt^2}{B^*}\right)+8\exp\left(-\frac{12^2nt^2}{B^*}\right)+8\exp\left(-\frac{12^2nt^2}{B^*}\right)+24\exp\left(-\frac{12^2nt^2}{B^*}\right)\right.\\
&\quad\left.+6\exp\left(-\frac{12^2nt^2}{B^*}\right)+6\exp\left(-\frac{12^2nt^2}{B^*}\right)+8\exp\left(-\frac{12^2nt^2}{B^*}\right)\right]\\
&=1-62\exp\left(-\frac{12^2nt^2}{B^*}\right),
\end{aligned}$$

where

$$\begin{aligned}
B^*:=\frac{1}{12^2}\max\big[&2B^4,8(BB_k,B_l)_+^4,8(BB_l,B_k)_+^4,18(B,B_k,B_l)_+^6,\\
&18(B_k^2,B_l)_+^6,18(B_k,B_l^2)_+^6,32\cdot 3^2(B_k,B_l)_+^8\big].
\end{aligned}$$

By reparameterization, it follows that

$$\mathbb{P}\left(\frac{c_1Jn}{\gamma_n}\left|\hat{S}(\mathbf{t},\mathbf{t}')-S(\mathbf{t},\mathbf{t}')\right|\le t\right)\ge 1-62\exp\left(-\frac{\gamma_n^2t^2}{c_1^2J^2nB^*}\right).\tag{4.22}$$

Union Bound for $\left|\hat{\lambda}_n-\lambda_n\right|$ and Final Lower Bound

We will bound terms in (4.21) separately and combine all the bounds with the union bound. As shown in (4.7), the U-statistic core h is bounded between $-2B$ and $2B$. Thus, by Lemma A.5 (with $m=2$), we have

$$\mathbb{P}\left(c_2n\sqrt{J}|\hat{u}(\mathbf{t}^*)-u(\mathbf{t}^*)|\le t\right)\ge 1-2\exp\left(-\frac{\lfloor 0.5n\rfloor t^2}{8c_2^2n^2JB^2}\right).\tag{4.23}$$

Bounding $\frac{c_1 n}{\gamma_n}\frac{8B^2J}{n-1} + c_3 n\gamma_n + \frac{4BJc_1 n}{\gamma_n}|\hat{u}(\tilde{\mathbf{t}}) - u(\tilde{\mathbf{t}})|$. By Lemma A.5 (with $m = 2$), it follows that

$$\begin{aligned}
&\mathbb{P}\left(\frac{c_1 n}{\gamma_n}\frac{8B^2J}{n-1} + c_3 n\gamma_n + \frac{4BJc_1 n}{\gamma_n}|\hat{u}(\tilde{\mathbf{t}}) - u(\tilde{\mathbf{t}})| \le t\right) \\
&\ge 1 - 2\exp\left(-\frac{\lfloor 0.5n\rfloor \gamma_n^2\left[t - \frac{c_1 n}{\gamma_n}\frac{8B^2J}{n-1} - c_3 n\gamma_n\right]^2}{2^7 B^4 J^2 c_1^2 n^2}\right) \\
&= 1 - 2\exp\left(-\frac{\lfloor 0.5n\rfloor \left[t\gamma_n(n-1) - 8c_1B^2nJ - c_3 n(n-1)\gamma_n^2\right]^2}{2^7 B^4 J^2 c_1^2 n^2 (n-1)^2}\right) \\
&\overset{(a)}{\ge} 1 - 2\exp\left(-\frac{\left[t\gamma_n(n-1) - 8c_1B^2nJ - c_3 n(n-1)\gamma_n^2\right]^2}{2^8 B^4 J^2 c_1^2 n^2 (n-1)}\right),
\end{aligned} \tag{4.24}$$

where at (a) we used $\lfloor 0.5n\rfloor \ge (n-1)/2$. Combining (4.22), (4.23), and (4.24) with the union bound (set $T = 3t$), we can bound (4.21) with

$$\begin{aligned}
\mathbb{P}\left(|\hat{\lambda}_n - \lambda_n| \le T\right) \ge 1 &- 62\exp\left(-\frac{\gamma_n^2 T^2}{3^2 c_1^2 J^2 n B^*}\right) - 2\exp\left(-\frac{\lfloor 0.5n\rfloor T^2}{72 c_2^2 n^2 J B^2}\right) \\
&- 2\exp\left(-\frac{\left[T\gamma_n(n-1)/3 - 8c_1B^2nJ - c_3\gamma_n^2 n(n-1)\right]^2}{2^8 B^4 J^2 c_1^2 n^2 (n-1)}\right).
\end{aligned}$$

Since $|\hat{\lambda}_n - \lambda_n| \le T$ implies $\hat{\lambda}_n \ge \lambda_n - T$, a reparametrization with $r = \lambda_n - T$ gives

$$\begin{aligned}
\mathbb{P}\left(\hat{\lambda}_n \ge r\right) &\ge 1 - 62\exp\left(-\frac{\gamma_n^2(\lambda_n - r)^2}{3^2 c_1^2 J^2 n B^*}\right) - 2\exp\left(-\frac{\lfloor 0.5n\rfloor(\lambda_n - r)^2}{72 c_2^2 n^2 J B^2}\right) \\
&\quad - 2\exp\left(-\frac{\left[(\lambda_n - r)\gamma_n(n-1)/3 - 8c_1B^2nJ - c_3\gamma_n^2 n(n-1)\right]^2}{2^8 B^4 J^2 c_1^2 n^2 (n-1)}\right) \\
&\quad := L(\lambda_n).
\end{aligned}$$

Grouping constants into $\xi_1, \dots \xi_5$ gives the result.

The lower bound $L(\lambda_n)$ takes the form

$$1 - 62\exp\left(-C_1(\lambda_n - T_\alpha)^2\right) - 2\exp\left(-C_2(\lambda_n - T_\alpha)^2\right) - 2\exp\left(-\frac{\left[(\lambda_n - T_\alpha)C_3 - C_4\right]^2}{C_5}\right),$$

where $C_1, \dots, C_5$ are positive constants. For fixed large enough n such that $\lambda_n > T_\alpha$, and fixed significance level α, increasing λ_n will increase $L(\lambda_n)$. Specifically, since n is fixed, increasing $\mathbf{u}^\top \boldsymbol{\Sigma}^{-1}\mathbf{u}$ in $\lambda_n = n\mathbf{u}^\top \boldsymbol{\Sigma}^{-1}\mathbf{u}$ will increase $L(\lambda_n)$.

4.B Helper Lemmas

This section contains contributed as well as known lemmas used to prove the main results in this chapter.

Lemma 4.7 (Product to sum)**.** *Assume that $|a_i| \le B$, $|b_i| \le B$ for $i = 1, \dots, E$. Then*

$\left|\prod_{i=1}^{E} a_i - \prod_{i=1}^{E} b_i\right| \leq B^{E-1} \sum_{j=1}^{E} |a_j - b_j|$.

Proof.

$$\begin{aligned}
\left|\prod_{i=1}^{E} a_i - \prod_{j=1}^{E} b_j\right| &\leq \left|\prod_{i=1}^{E} a_i - \prod_{i=1}^{E-1} a_i b_E\right| + \left|\prod_{i=1}^{E-1} a_i b_E - \prod_{i=1}^{E-2} a_i b_{E-1} b_E\right| + \ldots + \left|a_1 \prod_{j=2}^{E} b_j - \prod_{j=1}^{E} b_j\right| \\
&\leq |a_E - b_E| \left|\prod_{i=1}^{E-1} a_i\right| + |a_{E-1} - b_{E-1}| \left|\left(\prod_{i=1}^{E-2} a_i\right) b_E\right| + \ldots + |a_1 - b_1| \left|\prod_{j=2}^{E} b_j\right| \\
&\leq |a_E - b_E| B^{E-1} + |a_{E-1} - b_{E-1}| B^{E-1} + \ldots + |a_1 - b_1| B^{E-1} \\
&= B^{E-1} \sum_{j=1}^{E} |a_j - b_j|
\end{aligned}$$

applying triangle inequality, and the boundedness of a_i and b_i-s. □

Lemma 4.8 (Product variant of the Hoeffding's inequality)**.** *For $i = 1, \ldots, E$, let $\{\mathbf{x}_j^{(i)}\}_{j=1}^{n_i} \subset \mathcal{X}_i$ be an i.i.d. sample from a distribution P_i, and $f_i : \mathcal{X}_i \mapsto \mathbb{R}$ be a measurable function. Note that it is possible that $P_1 = P_2 = \cdots = P_E$ and $\{\mathbf{x}_j^{(1)}\}_{j=1}^{n_1} = \cdots = \{\mathbf{x}_j^{(E)}\}_{j=1}^{n_E}$. Assume that $|f_i(\mathbf{x})| \leq B < \infty$ for all $\mathbf{x} \in \mathcal{X}_i$ and $i = 1, \ldots, E$. Write $\hat{P}_i$ to denote an empirical distribution based on the sample $\{\mathbf{x}_j^{(i)}\}_{j=1}^{n_i}$. Then,*

$$\mathbb{P}\left(\left|\left[\prod_{i=1}^{E} \mathbb{E}_{\mathbf{x}^{(i)} \sim \hat{P}_i} f_i(\mathbf{x}^{(i)})\right] - \left[\prod_{i=1}^{E} \mathbb{E}_{\mathbf{x}^{(i)} \sim P_i} f_i(\mathbf{x}^{(i)})\right]\right| \leq T\right) \geq 1 - 2\sum_{i=1}^{E} \exp\left(-\frac{n_i T^2}{2E^2 B^{2E}}\right).$$

Proof. By Lemma 4.7, we have

$$\left|\left[\prod_{i=1}^{E} \mathbb{E}_{\mathbf{x}^{(i)} \sim \hat{P}_i} f_i(\mathbf{x}^{(i)})\right] - \left[\prod_{i=1}^{E} \mathbb{E}_{\mathbf{x}^{(i)} \sim P_i} f_i(\mathbf{x}^{(i)})\right]\right| \leq B^{E-1} \sum_{i=1}^{E} \left|\mathbb{E}_{\mathbf{x}^{(i)} \sim \hat{P}_i} f_i(\mathbf{x}^{(i)}) - \mathbb{E}_{\mathbf{x}^{(i)} \sim P_i} f_i(\mathbf{x}^{(i)})\right|.$$

By applying the Hoeffding's inequality to each term in the sum, we have

$$\mathbb{P}\left(\left|\mathbb{E}_{\mathbf{x}^{(i)} \sim \hat{P}_i} f_i(\mathbf{x}^{(i)}) - \mathbb{E}_{\mathbf{x}^{(i)} \sim P_i} f_i(\mathbf{x}^{(i)})\right| \leq t\right) \geq 1 - 2\exp\left(-\frac{2n_i t^2}{4B^2}\right).$$

The result is obtained with a union bound. □

Lemma 4.9 (Product of real analytic functions)**.** *Let $f\colon \mathcal{X} \to \mathbb{R}$ and $g\colon \mathcal{Y} \to \mathbb{R}$ be real analytic functions where $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ are open sets. Define $h(\mathbf{x}, \mathbf{y}) := f(\mathbf{x})g(\mathbf{y})$. Then, h is real analytic on $\mathcal{X} \times \mathcal{Y}$.*

Proof. By Lemma 4.10, f and g satisfy the following properties:

- For any $\mathbf{v} \in \mathcal{X}$, there exist an open ball W_f with $\mathbf{v} \in W_f \subseteq \mathcal{X}$, and constants $C_f > 0$ and $R_f > 0$ such that

$$\left|\frac{\partial^{\boldsymbol{\alpha}}}{\partial \mathbf{x}^{\boldsymbol{\alpha}}} f(\mathbf{x})\right| \leq C_f \frac{\boldsymbol{\alpha}!}{R_f^{|\boldsymbol{\alpha}|}}, \quad \forall \mathbf{x} \in W_f, \tag{4.25}$$

for any $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{d_x}) \in \{0, 1, 2, \ldots\}^{d_x}$.

- For any $\mathbf{w} \in \mathcal{Y}$, there exist an open ball W_g with $\mathbf{w} \in W_g \subseteq \mathcal{Y}$, and constants $C_g > 0$ and $R_g > 0$ such that

$$\left| \frac{\partial^{\boldsymbol{\beta}}}{\partial \mathbf{y}^{\boldsymbol{\beta}}} g(\mathbf{y}) \right| \leq C_g \frac{\boldsymbol{\beta}!}{R_g^{|\boldsymbol{\beta}|}}, \quad \forall \mathbf{y} \in W_g, \tag{4.26}$$

for any $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{d_y}) \in \{0, 1, 2, \ldots\}^{d_y}$.

Since f and g are infinitely differentiable, h is infinitely differentiable on $\mathcal{X} \times \mathcal{Y}$. (4.25) and (4.26) together mean that, for any $(\mathbf{v}, \mathbf{w}) \in \mathcal{X} \times \mathcal{Y}$, there exist an open ball $W \subset W_f \times W_g$ which contains $(\mathbf{v}, \mathbf{w})$, and constants $C := C_f C_g$ and $R := \min(R_f, R_g)$ such that

$$\left| \frac{\partial^{\boldsymbol{\alpha}} \partial^{\boldsymbol{\beta}}}{\partial \mathbf{x}^{\boldsymbol{\alpha}} \partial \mathbf{y}^{\boldsymbol{\beta}}} h(\mathbf{x}, \mathbf{y}) \right| \leq C \frac{\boldsymbol{\alpha}! \boldsymbol{\beta}!}{R^{|\boldsymbol{\alpha}| + |\boldsymbol{\beta}|}}, \quad \forall (\mathbf{x}, \mathbf{y}) \in W,$$

implying that $h(\mathbf{x}, \mathbf{y})$ is real analytic on $\mathcal{X} \times \mathcal{Y}$ by Lemma 4.10. □

Lemma 4.10 (Krantz and Parks [2002, Proposition 2.2.10])**.** *Let f be an infinitely differentiable function on an open set $\mathcal{X} \subseteq \mathbb{R}^d$. The function f is in fact real analytic if and only if, for each $\mathbf{v} \in \mathcal{X}$, there are an open ball W, with $\mathbf{v} \in W \subseteq \mathcal{X}$, and constants $C > 0$ and $R > 0$ such that the derivatives of f satisfy*

$$\left| \frac{\partial^{\mathbf{m}} f}{\partial \mathbf{x}^{\mathbf{m}}} (\mathbf{x}) \right| \leq C \frac{\mathbf{m}!}{R^{|\mathbf{m}|}}, \quad \forall \mathbf{x} \in W,$$

where $\mathbf{m} = (m_1, \ldots, m_d) \in \{0, 1, 2, \ldots\}^d$ is used as a multi-index with $\mathbf{x}^{\mathbf{m}} := \prod_{i=1}^d x_i^{m_i}$, $\mathbf{m}! := \prod_{i=1}^d m_i!$, $|\mathbf{m}| := \sum_{i=1}^d m_i$, and $\frac{\partial^{\mathbf{m}}}{\partial \mathbf{x}^{\mathbf{m}}} = \frac{\partial^{m_1}}{\partial x_1^{m_1}} \frac{\partial^{m_2}}{\partial x_2^{m_2}} \cdots \frac{\partial^{m_d}}{\partial x_d^{m_d}}$.

Chapter 5

Informative Features for Model Criticism

Summary We propose a novel adaptive test of goodness of fit, with computational cost linear in the number of samples. We learn the test features that best indicate the differences between observed samples and a reference model, by minimizing the false negative rate. These features are constructed via Stein's method, meaning that it is not necessary to compute the normalising constant of the model. We analyse the asymptotic Bahadur efficiency of the new test, and prove that under a mean-shift alternative, our test always has greater relative efficiency than a previous linear-time kernel test, regardless of the choice of parameters for that test. In experiments, the performance of our method exceeds that of the earlier linear-time test, and matches or exceeds the power of a quadratic-time kernel test. In high dimensions and where model structure may be exploited, our goodness of fit test performs far better than a quadratic-time two-sample test based on the Maximum Mean Discrepancy, with samples drawn from the model.

5.1 Introduction

The goal of goodness of fit testing is to determine how well a model density $p(\mathbf{x})$ fits an observed sample $\mathsf{D} = \{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X} \subseteq \mathbb{R}^d$ from an unknown distribution $q(\mathbf{x})$. This goal may be achieved via a hypothesis test, where the null hypothesis $H_0 \colon p = q$ is tested against $H_1 \colon p \neq q$. The problem of testing goodness of fit has a long history in statistics [Frank J. Massey, 1951], with a number of tests proposed for particular parametric models. Such tests can require space partitioning [Györfi and van der Meulen, 1990, Beirlant et al., 1994], which works poorly in high dimensions; or closed-form integrals under the model, which may be difficult to obtain, besides in certain special cases [Baringhaus and Henze, 1988, Bowman and Foster, 1993, Székely and Rizzo, 2005, Rizzo, 2009]. An alternative is to conduct a two-sample test using samples drawn from *both* p and q. This approach was taken by Lloyd and Ghahramani [2015], using a test based on the (quadratic-time) Maximum Mean Discrepancy [Gretton et al., 2012a], however this does not take advantage of the known structure of p (quite apart

from the increased computational cost of dealing with samples from p).

More recently, measures of discrepancy with respect to a model have been proposed based on Stein's method [Ley et al., 2017]. A Stein operator for p may be applied to a class of test functions, yielding functions that have zero expectation under p. Classes of test functions can include the $W^{2,\infty}$ Sobolev space [Gorham and Mackey, 2015], and reproducing kernel Hilbert spaces (RKHS) [Oates et al., 2017b]. Applications include variance reduction in Bayesian quadrature Oates et al. [2017b,a] and construction of variational density estimates. Statistical tests have been proposed by Chwialkowski et al. [2016], Liu et al. [2016] based on classes of Stein transformed RKHS functions, where the test statistic is the norm of the smoothness-constrained function with largest expectation under q . We will refer to this statistic as the Kernel Stein Discrepancy (KSD). For consistent tests, it is sufficient to use c_0-universal kernels [Carmeli et al., 2010, Definition 4.1], as shown by Chwialkowski et al. [2016, Theorem 2.2], although inverse multiquadric kernels may be preferred if uniform tightness is required [Gorham and Mackey, 2017].[1]

The minimum variance unbiased estimate of the KSD is a U-statistic, with computational cost quadratic in the number n of samples from q. It is desirable to reduce the cost of testing, however, so that larger sample sizes may be addressed. A first approach is to replace the U-statistic with a running average with linear cost, as proposed by Liu et al. [2016] for the KSD, but this results in an increase in variance and corresponding decrease in test power. An alternative approach is to construct explicit features of the distributions, whose empirical expectations may be computed in linear time. In the two-sample and independence settings, these features were initially chosen at random by [Epps and Singleton, 1986, Chwialkowski et al., 2015, Zhang et al., 2017]. More recently, features have been constructed explicitly to maximize test power in the two-sample [Jitkrittum et al., 2016] and independence testing [Jitkrittum et al., 2017] settings, resulting in tests that are not only more interpretable, but which can yield performance matching quadratic-time tests.

We propose to construct explicit linear-time features for testing goodness of fit, chosen so as to maximize test power. These features further reveal where the model and data differ, in a readily interpretable way. Our first theoretical contribution is a derivation of the null and alternative distributions for tests based on such features, and a corresponding power optimization criterion. Note that the goodness-of-fit test requires somewhat different strategies to those employed for two-sample (Section 3.2) and independence testing (Section 4.3), which become computationally prohibitive in high dimensions for the Stein discrepancy (specifically, the normalization used in prior work to simplify the asymptotics would incur a cost cubic in the dimension d and the number of features in the optimization).

[1] Briefly, Gorham and Mackey show that when an exponentiated quadratic kernel is used, a sequence of sets D may be constructed that does not correspond to any q, but for which the KSD nonetheless approaches zero. In a statistical testing setting, however, we assume identically distributed samples from q, and the issue does not arise.

Our second theoretical contribution, given in Section 5.4, is an analysis of the relative Bahadur efficiency of our test vs the linear time test of Liu et al. [2016]: this represents the relative rate at which the p-value decreases under H_1 as we observe more samples. We prove that our test has greater asymptotic Bahadur efficiency relative to the test of Liu et al., for Gaussian distributions under the mean-shift alternative. This is shown to hold regardless of the bandwidth of the exponentiated quadratic kernel used for the earlier test. The proof techniques developed are of independent interest, and we anticipate that they may provide a foundation for the analysis of relative efficiency of linear-time tests in the two-sample and independence testing domains. In experiments (Section 5.5), our new linear-time test is able to detect subtle local differences between the density $p(\mathbf{x})$, and the unknown $q(\mathbf{x})$ as observed through samples. We show that our linear-time test constructed based on optimized features has comparable performance to the quadratic-time test of Chwialkowski et al. [2016], Liu et al. [2016], while uniquely providing an explicit visual indication of where the model fails to fit the data.

5.2 Kernel Stein Discrepancy (KSD) Test

We begin by introducing the Kernel Stein Discrepancy (KSD) and associated statistical test, as proposed independently by Chwialkowski et al. [2016] and Liu et al. [2016]. Assume that the data domain is a connected open set $\mathcal{X} \subseteq \mathbb{R}^d$. Consider a Stein operator T_p that takes in a multivariate function $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_d(\mathbf{x}))^\top \in \mathbb{R}^d$ and constructs a function $\left(T_p\mathbf{f}\right)(\mathbf{x}) \colon \mathbb{R}^d \to \mathbb{R}$. The constructed function has the key property that $\mathbb{E}_{\mathbf{x}\sim q}\left[(T_p\mathbf{f})(\mathbf{x})\right] = 0$ if and only if $q = p$, for all $\mathbf{f}$ in an appropriate function class. Thus, one can use this expectation as a statistic for testing goodness of fit.

The function class $\mathcal{F}^d$ for the function $\mathbf{f}$ is chosen to be a unit-norm ball in a reproducing kernel Hilbert space (RKHS) in Chwialkowski et al. [2016], Liu et al. [2016]. More precisely, let $\mathcal{F}$ be an RKHS associated with a positive definite kernel $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let $\phi(\mathbf{x}) = k(\mathbf{x}, \cdot)$ denote a feature map of k so that $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{F}}$. Assume that $f_i \in \mathcal{F}$ for all $i = 1, \ldots, d$ so that $\mathbf{f} \in \mathcal{F} \times \cdots \times \mathcal{F} := \mathcal{F}^d$ where $\mathcal{F}^d$ is equipped with the standard inner product $\langle \mathbf{f}, \mathbf{g} \rangle_{\mathcal{F}^d} := \sum_{i=1}^d \langle f_i, g_i \rangle_{\mathcal{F}}$. The kernelized Stein operator[2] T_p studied in Chwialkowski et al. [2016] is

$$\left(T_p\mathbf{f}\right)(\mathbf{x}) := \sum_{i=1}^{d} \left(\frac{\partial \log p(\mathbf{x})}{\partial x_i} f_i(\mathbf{x}) + \frac{\partial f_i(\mathbf{x})}{\partial x_i} \right) \overset{(a)}{=} \left\langle \mathbf{f}, \boldsymbol{\xi}_p(\mathbf{x}, \cdot) \right\rangle_{\mathcal{F}^d},$$

where at (a) we use the reproducing property of $\mathcal{F}$, i.e., $f_i(\mathbf{x}) = \langle f_i, k(\mathbf{x}, \cdot) \rangle_{\mathcal{F}}$, and that $\frac{\partial k(\mathbf{x}, \cdot)}{\partial x_i} \in \mathcal{F}$ [Steinwart and Christmann, 2008, Lemma 4.34], hence $\boldsymbol{\xi}_p(\mathbf{x}, \cdot) := \frac{\partial \log p(\mathbf{x})}{\partial \mathbf{x}} k(\mathbf{x}, \cdot) + \frac{\partial k(\mathbf{x}, \cdot)}{\partial \mathbf{x}}$ is in $\mathcal{F}^d$. Under appropriate conditions, e.g. that $\lim_{\|\mathbf{x}\| \to \infty} p(\mathbf{x}) f_i(\mathbf{x}) = 0$ for all $i = 1, \ldots, d$, it can be shown using integration by parts that $\mathbb{E}_{\mathbf{x}\sim p}(T_p\mathbf{f})(\mathbf{x}) = 0$

[2]The Stein operator presented in Liu et al. [2016] is defined such that $\left(T_p\mathbf{f}\right)(\mathbf{x}) \in \mathbb{R}^d$. This distinction is not crucial and leads to the same goodness-of-fit test.

for any $\mathbf{f} \in \mathcal{F}^d$ [Chwialkowski et al., 2016, Lemma 5.1]. Based on the Stein operator, Chwialkowski et al. [2016], Liu et al. [2016] define the kernelized Stein discrepancy as

$$S_p(q) := \sup_{\|\mathbf{f}\|_{\mathcal{F}^d} \leq 1} \mathbb{E}_{\mathbf{x}\sim q} \left\langle \mathbf{f}, \boldsymbol{\xi}_p(\mathbf{x}, \cdot) \right\rangle_{\mathcal{F}^d} \overset{(a)}{=} \sup_{\|\mathbf{f}\|_{\mathcal{F}^d} \leq 1} \left\langle \mathbf{f}, \mathbb{E}_{\mathbf{x}\sim q} \boldsymbol{\xi}_p(\mathbf{x}, \cdot) \right\rangle_{\mathcal{F}^d} = \|\mathbf{g}(\cdot)\|_{\mathcal{F}^d}, \quad (5.1)$$

where at (a), $\boldsymbol{\xi}_p(\mathbf{x}, \cdot)$ is Bochner integrable [Steinwart and Christmann, 2008, Definition A.5.20] as long as $\mathbb{E}_{\mathbf{x}\sim q} \|\boldsymbol{\xi}_p(\mathbf{x}, \cdot)\|_{\mathcal{F}^d} < \infty$, and $\mathbf{g}(\mathbf{y}) := \mathbb{E}_{\mathbf{x}\sim q} \boldsymbol{\xi}_p(\mathbf{x}, \mathbf{y})$ is what we refer to as the *Stein witness function*. The Stein witness function will play a crucial role in our new test statistic in Section 5.3. It can be shown that $S_p(q) = 0$ if and only if $p = q$ under some conditions (Theorem 5.1).

Theorem 5.1 (Chwialkowski et al. [2016, Theorem 2.2]). *If the kernel k is c_0-universal [Carmeli et al., 2010, Definition 4.1], $\mathbb{E}_{\mathbf{x}\sim q} \mathbb{E}_{\mathbf{x}'\sim q} h_p(\mathbf{x}, \mathbf{x}') < \infty$, and $\mathbb{E}_{\mathbf{x}\sim q} \|\nabla_{\mathbf{x}} \log \frac{p(\mathbf{x})}{q(\mathbf{x})}\|^2 < \infty$, then $S_p(q) = \|\mathbb{E}_{\mathbf{x}\sim q} \xi_p(\mathbf{x}, \cdot)\|_{\mathcal{F}^d} = 0$ if and only if $p = q$.*

The KSD $S_p(q)$ can be written as $S_p^2(q) = \mathbb{E}_{\mathbf{x}\sim q} \mathbb{E}_{\mathbf{x}'\sim q} h_p(\mathbf{x}, \mathbf{x}')$, where

$$h_p(\mathbf{x}, \mathbf{y}) := \mathbf{s}_p^\top(\mathbf{x}) \mathbf{s}_p(\mathbf{y}) k(\mathbf{x}, \mathbf{y}) + \mathbf{s}_p^\top(\mathbf{y}) \nabla_{\mathbf{x}} k(\mathbf{x}, \mathbf{y}) + \mathbf{s}_p^\top(\mathbf{x}) \nabla_{\mathbf{y}} k(\mathbf{x}, \mathbf{y}) + \sum_{i=1}^{d} \frac{\partial^2 k(\mathbf{x}, \mathbf{y})}{\partial x_i \partial y_i}, \quad (5.2)$$

and $\mathbf{s}_p(\mathbf{x}) := \nabla_{\mathbf{x}} \log p(\mathbf{x})$ is a column vector. An unbiased empirical estimator of $S_p^2(q)$, denoted by $\widehat{S^2} = \frac{2}{n(n-1)} \sum_{i<j} h_p(\mathbf{x}_i, \mathbf{x}_j)$ [Liu et al., 2016, Eq. 14], is a degenerate U-statistic under H_0. For the goodness-of-fit test, the rejection threshold can be computed by a bootstrap procedure. All these properties make $\widehat{S^2}$ a very flexible criterion to detect the discrepancy of p and q: in particular, it can be computed even if p is known only up to a normalization constant. Further studies on nonparametric Stein operators can be found in Oates et al. [2017b], Gorham and Mackey [2015].

Linear-Time Kernel Stein (LKS) Test Computation of $\widehat{S^2}$ costs $\mathcal{O}(n^2)$. To reduce this cost, a linear-time (i.e., $\mathcal{O}(n)$) estimator based on an incomplete U-statistic is proposed in Liu et al. [2016, Eq. 17], given by

$$\widehat{S_l^2} := \frac{2}{n} \sum_{i=1}^{n/2} h_p(\mathbf{x}_{2i-1}, \mathbf{x}_{2i}), \quad (5.3)$$

where we assume n is even for simplicity. Empirically Liu et al. [2016] observed that the linear-time estimator performs much worse (in terms of test power) than the quadratic-time U-statistic estimator, agreeing with our findings presented in Section 5.5.

5.3 New Statistic: The Finite Set Stein Discrepancy (FSSD)

Although shown to be powerful, the main drawback of the Kernel Stein Discrepancy test is its high computational cost of $\mathcal{O}(n^2)$. Further, to get the rejection threshold, if the bootstrap resampling is performed b times, then the complexity is $\mathcal{O}(bn^2)$ which

is high, making it impractical for data with large sample size. The LKS test is one order of magnitude faster. Unfortunately, the decrease in the test power outweighs the computational gain [Liu et al., 2016]. We therefore seek a variant of the KSD statistic that can be computed in linear time, and whose test power is comparable to the KSD test.

Key Idea The fact that $S_p(q) = 0$ if and only if $p = q$ implies that $\mathbf{g}(\mathbf{v}) = \mathbf{0}$ for all $\mathbf{v} \in \mathcal{X}$ if and only if $p = q$, where $\mathbf{g}$ is the Stein witness function in (5.1). One can see $\mathbf{g}$ as a function witnessing the differences of p, q, in such a way that $|g_i(\mathbf{v})|$ is large when there is a discrepancy in the region around $\mathbf{v}$, as indicated by the i^{th} output of $\mathbf{g}$. The test statistic of Liu et al. [2016], Chwialkowski et al. [2016] is essentially given by the degree of "flatness" of $\mathbf{g}$ as measured by the RKHS norm $\|\cdot\|_{\mathcal{F}^d}$. The core of our proposal is to use a different measure of flatness of $\mathbf{g}$ which can be computed in linear time.

The idea is to use a real analytic kernel k which makes $g_1, \ldots, g_d$ real analytic. If $g_i \neq 0$ is an analytic function, then the Lebesgue measure of the set of roots $\{\mathbf{x} \mid g_i(\mathbf{x}) = 0\}$ is zero [Mityagin, 2015]. This property suggests that one can evaluate g_i at a finite set of locations $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$, drawn from a distribution with a density (w.r.t. the Lebesgue measure). If $g_i \neq 0$, then almost surely $g_i(\mathbf{v}_1), \ldots, g_i(\mathbf{v}_J)$ will not be zero. Our new test statistic based on this idea is called the Finite Set Stein Discrepancy (FSSD) and is given in Theorem 5.2.

Theorem 5.2 (The Finite Set Stein Discrepancy (FSSD))**.** *Let $\mathcal{X}$ be a connected open set in $\mathbb{R}^d$. Define $\mathrm{FSSD}_p^2(q) := \frac{1}{dJ}\sum_{i=1}^{d}\sum_{j=1}^{J} g_i^2(\mathbf{v}_j)$. Assume*

1. *The test locations $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\} \subset \mathbb{R}^d$ are drawn i.i.d. from a distribution η which has a density,*
2. *$k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is c_0-universal and real analytic (see Section 2.5),*
3. *$\mathbb{E}_{\mathbf{x}\sim q}\mathbb{E}_{\mathbf{x}'\sim q} h_p(\mathbf{x}, \mathbf{x}') < \infty$,*
4. *$\mathbb{E}_{\mathbf{x}\sim q}\|\nabla_{\mathbf{x}} \log p(\mathbf{x}) - \nabla_{\mathbf{x}} \log q(\mathbf{x})\|^2 < \infty$, and*
5. *$\lim_{\|\mathbf{x}\|\to\infty} p(\mathbf{x})\mathbf{g}(\mathbf{x}) = 0$.*

Then, η-almost surely $\mathrm{FSSD}_p^2(q) = 0$ if and only if $p = q$, for any $J \geq 1$.

Proof. Since k is real analytic, the components $g_1, \ldots, g_d$ of $\mathbf{g}$ are real analytic by Lemma 2.12. For each $i = 1, \ldots, d$, if g_i is real analytic, then $\sum_{j=1}^{J} g_i^2(\mathbf{v}_j) = 0$ if and only if $g_i(\mathbf{y}) = 0$ for all $\mathbf{y} \in \mathcal{X}$, η-almost surely [Mityagin, 2015]. This implies that $\frac{1}{dJ}\sum_{i=1}^{d}\sum_{j=1}^{J} g_i^2(\mathbf{v}_j) = 0$ if and only if $\mathbf{g}(\mathbf{y}) = \mathbf{0}$ for all $\mathbf{y} \in \mathcal{X}$, η-almost surely. By Theorem 5.1, $\mathbf{g} = \mathbf{0}$ (the zero function) if and only if $p = q$. □

This measure depends on a set of J test locations (or features) $\{\mathbf{v}_i\}_{i=1}^{J}$ used to evaluate the Stein witness function, where J is fixed and is typically small. A kernel which is c_0-universal and real analytic is the Gaussian kernel $k(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2\sigma_k^2}\right)$. We will consider only the Gaussian kernel. Besides the requirement that the kernel be

real and analytic, the remaining conditions in Theorem 5.2 are the same as given in Chwialkowski et al. [2016, Theorem 2.2]. Note that if the FSSD is to be employed in a setting otherwise than testing, for instance to obtain pseudo-samples converging to p, then stronger conditions may be needed [Gorham and Mackey, 2017].

5.3.1 Goodness-of-Fit Test with the FSSD Statistic

Given a significance level α for the goodness-of-fit test, the test can be constructed so that H_0 is rejected when $n\widehat{\text{FSSD}^2} > T_\alpha$, where T_α is the rejection threshold (critical value), and $\widehat{\text{FSSD}^2}$ is an empirical estimate of $\text{FSSD}_p^2(q)$. The threshold which guarantees that the type-I error (i.e., the probability of rejecting H_0 when it is true) is bounded above by α is given by the $(1-\alpha)$-quantile of the null distribution i.e., the distribution of $n\widehat{\text{FSSD}^2}$ under H_0. In the following, we start by giving the expression for $\widehat{\text{FSSD}^2}$, and summarize its asymptotic distributions in Proposition 5.3.

Let $\Xi(\mathbf{x}) \in \mathbb{R}^{d\times J}$ such that $[\Xi(\mathbf{x})]_{i,j} = \xi_{p,i}(\mathbf{x}, \mathbf{v}_j)/\sqrt{dJ}$. Define $\boldsymbol{\tau}(\mathbf{x}) := \text{vec}(\Xi(\mathbf{x})) \in \mathbb{R}^{dJ}$ where $\text{vec}(\mathbf{M})$ concatenates columns of the matrix $\mathbf{M}$ into a column vector. We note that $\boldsymbol{\tau}(\mathbf{x})$ depends on the test locations $V = \{\mathbf{v}_j\}_{j=1}^J$. Let $\Delta(\mathbf{x}, \mathbf{y}) := \boldsymbol{\tau}(\mathbf{x})^\top \boldsymbol{\tau}(\mathbf{y}) = \text{tr}(\Xi(\mathbf{x})^\top \Xi(\mathbf{y}))$. Given an i.i.d. sample $\{\mathbf{x}_i\}_{i=1}^n \sim q(\mathbf{x})$, a consistent, unbiased estimator of $\text{FSSD}_p^2(q)$ is

$$\begin{aligned}\widehat{\text{FSSD}^2} &= \frac{1}{dJ}\sum_{l=1}^{d}\sum_{m=1}^{J}\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\xi_{p,l}(\mathbf{x}_i, \mathbf{v}_m)\xi_{p,l}(\mathbf{x}_j, \mathbf{v}_m) \\ &= \frac{2}{n(n-1)}\sum_{i<j}\Delta(\mathbf{x}_i, \mathbf{x}_j),\end{aligned} \tag{5.4}$$

which is a one-sample second-order U-statistic with Δ as its U-statistic core (see Section A.1: U-Statistics). Being a U-statistic, its asymptotic distribution can easily be derived. We use $\xrightarrow{d}$ to denote convergence in distribution.

Proposition 5.3 (Asymptotic distributions of $\widehat{\text{FSSD}^2}$)**.** *Let $Z_1, \ldots, Z_{dJ} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Let $\boldsymbol{\mu} := \mathbb{E}_{\mathbf{x}\sim q}[\boldsymbol{\tau}(\mathbf{x})]$, $\boldsymbol{\Sigma}_r := \text{cov}_{\mathbf{x}\sim r}[\boldsymbol{\tau}(\mathbf{x})] \in \mathbb{R}^{dJ\times dJ}$ for $r \in \{p, q\}$, and $\{\omega_i\}_{i=1}^{dJ}$ be the eigenvalues of $\boldsymbol{\Sigma}_p = \mathbb{E}_{\mathbf{x}\sim p}[\boldsymbol{\tau}(\mathbf{x})\boldsymbol{\tau}^\top(\mathbf{x})]$. Assume that $\mathbb{E}_{\mathbf{x}\sim q}\mathbb{E}_{\mathbf{y}\sim q}\Delta^2(\mathbf{x}, \mathbf{y}) < \infty$, and the assumptions in Theorem 5.2 hold. Then,*

1. *Under $H_0 : p = q$, $n\widehat{\text{FSSD}^2} \xrightarrow{d} \sum_{i=1}^{dJ}(Z_i^2 - 1)\omega_i$.*
2. *Under $H_1 : p \neq q$, if $\sigma_{H_1}^2 := 4\boldsymbol{\mu}^\top \boldsymbol{\Sigma}_q \boldsymbol{\mu} > 0$, then $\sqrt{n}(\widehat{\text{FSSD}^2} - \text{FSSD}^2) \xrightarrow{d} \mathcal{N}(0, \sigma_{H_1}^2)$.*

Proof. Recognizing that (5.4) is a degenerate U-statistic, the results follow directly from Serfling [2009, Section 5.5.1, 5.5.2]. □

Claims 1 and 2 of Proposition 5.3 imply that under H_1, the test power (i.e., the probability of correctly rejecting H_1) goes to 1 asymptotically, if the threshold T_α is defined as above. In practice, simulating from the asymptotic null distribution in Claim 1 can be challenging, since the plug-in estimator of $\boldsymbol{\Sigma}_p$ requires a sample from p, which is not available. A straightforward solution is to draw sample from p, either

by assuming that p can be sampled easily or by using an MCMC method, although this adds an additional computational burden to the test procedure. A more subtle issue is that when dependent samples from p are used in obtaining the test threshold, the test may become more conservative than required for i.i.d. data [Chwialkowski et al., 2014]. An alternative approach is to use the plug-in estimate $\hat{\boldsymbol{\Sigma}}_q$ instead of $\boldsymbol{\Sigma}_p$. The covariance matrix $\hat{\boldsymbol{\Sigma}}_q$ can be directly computed from the data. This is the approach we take. Theorem 5.4 guarantees that the replacement of the covariance in the computation of the asymptotic null distribution still yields a consistent test. We write $\mathbb{P}_{H_1}$ for the distribution of $n\widehat{\mathrm{FSSD}^2}$ under H_1.

Theorem 5.4. *Let* $\hat{\boldsymbol{\Sigma}}_q := \frac{1}{n}\sum_{i=1}^n \boldsymbol{\tau}(\mathbf{x}_i)\boldsymbol{\tau}^\top(\mathbf{x}_i) - [\frac{1}{n}\sum_{i=1}^n \boldsymbol{\tau}(\mathbf{x}_i)][\frac{1}{n}\sum_{j=1}^n \boldsymbol{\tau}(\mathbf{x}_j)]^\top$ *with* $\{\mathbf{x}_i\}_{i=1}^n \sim q$. *Suppose that the test threshold* T_α *is set to the* $(1-\alpha)$*-quantile of the distribution of* $\sum_{i=1}^{dJ} \hat{\nu}_i(Z_i^2 - 1)$ *where* $\{Z_i\}_{i=1}^{dJ} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, *and* $\hat{\nu}_1, \ldots, \hat{\nu}_{dJ}$ *are eigenvalues of* $\hat{\boldsymbol{\Sigma}}_q$. *Assume that the assumptions in Theorem 5.2 hold. Then, under* H_0, *asymptotically the false rejection rate is* α. *Under* H_1, *the test power* $\mathbb{P}_{H_1}(n\widehat{\mathrm{FSSD}^2} > T_\alpha) \to 1$ *as* $n \to \infty$.

Proof. Under H_0, $p = q$ implies that $\hat{\boldsymbol{\Sigma}}_q = \hat{\boldsymbol{\Sigma}}_p$ (empirical estimate of $\boldsymbol{\Sigma}_p$). Let $\lambda_j(A)$ denote the j^{th} eigenvalue of the matrix A. Lemma 5.15 implies that $A \mapsto \lambda_j(A)$ is continuous on the space of real symmetric matrices, for all j. Since $\mathrm{plim}_{n\to\infty} \|\hat{\boldsymbol{\Sigma}}_p - \boldsymbol{\Sigma}_p\| = 0$ (i.e., $\hat{\boldsymbol{\Sigma}}_p$ converges in probability to $\boldsymbol{\Sigma}_p$), by the continuous mapping theorem, the eigenvalues of $\hat{\boldsymbol{\Sigma}}_p$ converge to the eigenvalues of $\boldsymbol{\Sigma}_p$ in probability. This implies that $\sum_{i=1}^{dJ}(Z_i^2 - 1)\hat{\nu}_i$ converges in probability to $\sum_{i=1}^{dJ}(Z_i^2 - 1)\omega_i$ as $n \to \infty$, where $\{\omega_i\}_{i=1}^{dJ}$ are eigenvalues of $\boldsymbol{\Sigma}_p$. By Lemma 5.16, the quantile also converges, and the test threshold thus matches that of the true asymptotic null distribution given in claim 1 of Proposition 5.3.

Assume H_1 holds. Let $\hat{t}_\alpha, t_\alpha$ be $(1-\alpha)$-quantiles of the distributions of $\sum_{i=1}^{dJ}(Z_i^2 - 1)\hat{\nu}_i$ and $\sum_{i=1}^{dJ}(Z_i^2 - 1)\nu_i$, respectively, where $\{\nu_i\}_{i=1}^{dJ}$ are eigenvalues of $\boldsymbol{\Sigma}_q$. By the same argument as in the previous paragraph, $\hat{t}_\alpha$ converges in probability to t_α, which is a constant independent of the sample size n. Given $\{\mathbf{v}_j\}_{j=1}^J \sim \eta$, where η is a distribution with a density, $\mathrm{FSSD}^2 > 0$ by Theorem 5.2. It follows that

$$\lim_{n\to\infty} \mathbb{P}\left(n\widehat{\mathrm{FSSD}^2} > \hat{t}_\alpha\right) = \lim_{n\to\infty} \mathbb{P}\left(\widehat{\mathrm{FSSD}^2} - \frac{\hat{t}_\alpha}{n} > 0\right) \overset{(a)}{=} \mathbb{P}\left(\mathrm{FSSD}^2 > 0\right) = 1,$$

where at (a), we use the fact that $\widehat{\mathrm{FSSD}^2}$ converges in probability to FSSD^2 by the law of large numbers, and that $\hat{t}_\alpha / n \xrightarrow{p} 0$. □

5.3.2 Optimizing the Test Parameters

Theorem 5.2 guarantees that the population quantity $\mathrm{FSSD}^2 = 0$ if and only if $p = q$ for any choice of $\{\mathbf{v}_i\}_{i=1}^J$ drawn from a distribution with a density. In practice, we are forced to rely on the empirical $\widehat{\mathrm{FSSD}^2}$, and some test locations will give a higher detection rate (i.e., test power) than others for finite n. Following the approaches of Gretton et al. [2012b], Jitkrittum et al. [2016], Sutherland et al. [2016], Jitkrittum et al. [2017], we choose the test locations $V = \{\mathbf{v}_j\}_{j=1}^J$ and kernel bandwidth σ_k^2 so as to

maximize the test power i.e., the probability of rejecting H_0 when it is false. We first give an approximate expression for the test power when n is large.

Proposition 5.5 (Approximate test power of $n\widehat{\text{FSSD}^2}$)**.** *Under H_1, for large n and fixed r, the test power $\mathbb{P}_{H_1}(n\widehat{\text{FSSD}^2} > r) \approx 1 - \Phi\left(\frac{r}{\sqrt{n}\sigma_{H_1}} - \sqrt{n}\frac{\text{FSSD}^2}{\sigma_{H_1}}\right)$, where Φ denotes the cumulative distribution function of the standard normal distribution, and σ_{H_1} is defined in Proposition 5.3.*

Proof. $\mathbb{P}_{H_1}(n\widehat{\text{FSSD}^2} > r) = \mathbb{P}_{H_1}(\widehat{\text{FSSD}^2} > r/n) = \mathbb{P}_{H_1}\left(\sqrt{n}\frac{\widehat{\text{FSSD}^2} - \text{FSSD}^2}{\sigma_{H_1}} > \sqrt{n}\frac{r/n - \text{FSSD}^2}{\sigma_{H_1}}\right)$. For sufficiently large n, the alternative distribution is approximately normal as given in Proposition 5.3. It follows that $\mathbb{P}_{H_1}(n\widehat{\text{FSSD}^2} > r) \approx 1 - \Phi\left(\frac{r}{\sqrt{n}\sigma_{H_1}} - \sqrt{n}\frac{\text{FSSD}^2}{\sigma_{H_1}}\right)$. □

Let $\zeta := \{V, \sigma_k^2\}$ be the collection of all tuning parameters. Assume that n is sufficiently large. Following the same argument as in Sutherland et al. [2016], in $\frac{r}{\sqrt{n}\sigma_{H_1}} - \sqrt{n}\frac{\text{FSSD}^2}{\sigma_{H_1}}$, we observe that the first term $\frac{r}{\sqrt{n}\sigma_{H_1}} = \mathcal{O}(n^{-1/2})$ going to 0 as $n \to \infty$, while the second term $\sqrt{n}\frac{\text{FSSD}^2}{\sigma_{H_1}} = \mathcal{O}(n^{1/2})$, dominating the first for large n. Thus, the best parameters that maximize the test power are given by

$$\zeta^* = \arg\max_{\zeta} \mathbb{P}_{H_1}(n\widehat{\text{FSSD}^2} > T_\alpha) \approx \arg\max_{\zeta} \frac{\text{FSSD}^2}{\sigma_{H_1}}. \tag{5.5}$$

Since FSSD^2 and σ_{H_1} are unknown, we divide the sample $\{\mathbf{x}_i\}_{i=1}^n$ into two disjoint training and test sets, and use the training set to compute $\frac{\widehat{\text{FSSD}^2}}{\hat{\sigma}_{H_1}+\gamma}$ to estimate $\frac{\text{FSSD}^2}{\sigma_{H_1}}$, where a small regularization parameter $\gamma > 0$ is added for numerical stability. The goodness-of-fit test is performed on the test set to avoid overfitting. The idea of splitting the data into training and test sets to learn good features for hypothesis testing was successfully used in Sutherland et al. [2016], Jitkrittum et al. [2016, 2017], Gretton et al. [2012b].

To find a local maximum of $\frac{\widehat{\text{FSSD}^2}}{\hat{\sigma}_{H_1}+\gamma}$, we use gradient ascent for its simplicity. The initial points of $\{\mathbf{v}_i\}_{i=1}^J$ are set to random draws from a normal distribution fitted to the training data, a heuristic we found to perform well in practice. The objective is non-convex in general, reflecting many possible ways to capture the differences of p and q. The regularization parameter γ is not tuned, and is fixed to a small constant. Assume that $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ costs $\mathcal{O}(d^2)$ to evaluate. Computing $\nabla_\zeta \frac{\widehat{\text{FSSD}^2}}{\hat{\sigma}_{H_1}+\gamma}$ costs $\mathcal{O}(d^2J^2n)$. The computational complexity of estimating $n\widehat{\text{FSSD}^2}$ and $\hat{\sigma}_{H_1}^2$ is $\mathcal{O}(d^2Jn)$. Thus, finding a local optimum via gradient ascent is still linear-time in n, for a fixed maximum number of iterations. Computing $\hat{\boldsymbol{\Sigma}}_q$ costs $\mathcal{O}(d^2J^2n)$, and obtaining all the eigenvalues of $\hat{\boldsymbol{\Sigma}}_q$ costs $\mathcal{O}(d^3J^3)$. We note that the optimization does not require eigenvalues of $\hat{\boldsymbol{\Sigma}}_q$. The eigenvalues are only needed once to perform the test. If the eigenvalues decay to zero sufficiently rapidly, one can approximate the asymptotic null distribution with only a few eigenvalues. The cost to obtain the largest few eigenvalues alone can be much smaller.

Remark 1. Let $\hat{\boldsymbol{\mu}} := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\tau}(\mathbf{x}_i)$. It is possible to normalize the FSSD statistic to get a new statistic $\hat{\lambda}_n := n\hat{\boldsymbol{\mu}}^\top (\hat{\boldsymbol{\Sigma}}_q + \gamma \mathbf{I})^{-1}\hat{\boldsymbol{\mu}}$ where $\gamma \geq 0$ is a regularization parameter that goes to 0 as $n \to \infty$. This was done in the case of the normalized ME (mean embeddings) statistic in Section 3.2, and the normalized FSIC statistic in Section 4.3. The asymptotic null distribution of the normalized statistic in this case takes the convenient form of $\chi^2(dJ)$ (independent of p and q), eliminating the need to obtain the eigenvalues of $\hat{\boldsymbol{\Sigma}}_q$. As in the case of the normalized ME and FSIC statistics, the test power criterion for tuning the parameters in this case is the statistic $\hat{\lambda}_n$ itself. However, the optimization is computationally expensive as $(\hat{\boldsymbol{\Sigma}}_q + \gamma \mathbf{I})^{-1}$ (costing $\mathcal{O}(d^3J^3)$) needs to be reevaluated in each gradient ascent iteration. In particular, the cost is high even for a moderate value of the input dimension d. In our proposed FSSD statistic, there is no cubic dependency on d or J in the optimization.

5.4 Relative Efficiency of the FSSD and LKS Tests

Both the linear-time kernel Stein (LKS) and FSSD tests have the same computational cost of $\mathcal{O}(d^2 n)$, and are consistent, achieving maximum test power of 1 as $n \to \infty$ under H_1. It is thus of theoretical interest to understand which test is more sensitive in detecting the differences of p and q. This can be quantified by the *Bahadur slope* of the test [Bahadur, 1960]. Two given tests can then be compared by computing the *Bahadur efficiency* (Theorem 5.14) which is given by the ratio of the slopes of the two tests. We note that the constructions and techniques in this section may be of independent interest, and can be generalised to other statistical testing settings.

5.4.1 Relative Efficiency and Bahadur Slope

We start by introducing the concept of Bahadur slope for a general test, following the presentation of Gleser [1964, 1966]. Consider a hypothesis testing problem on a parameter θ. The test proposes a null hypothesis $H_0 : \theta \in \Theta_0$ against the alternative hypothesis $H_1 : \theta \in \Theta \backslash \Theta_0$, where Θ, Θ_0 are arbitrary sets. Let T_n be a test statistic computed from a sample of size n, such that large values of T_n provide an evidence to reject H_0. We use plim to denote convergence in probability, and write $\mathbb{E}_r$ for $\mathbb{E}_{\mathbf{x}\sim r}\mathbb{E}_{\mathbf{x}'\sim r}$.

Approximate Bahadur Slope (ABS) For $\theta_0 \in \Theta_0$, let the asymptotic null distribution of T_n be $F(t) = \lim_{n\to\infty} P_{\theta_0}(T_n < t)$, where we assume that the CDF (F) is continuous and common to all $\theta_0 \in \Theta_0$. The continuity of F will be important later when Theorems 5.7 and 5.8 are used to compute the slopes of LKS and FSSD tests. Assume that there exists a continuous strictly increasing function $\rho : (0,\infty) \to (0,\infty)$ such that $\lim_{n\to\infty} \rho(n) = \infty$, and that $-2\,\mathrm{plim}_{n\to\infty} \frac{\log(1-F(T_n))}{\rho(n)} = c(\theta)$ where $T_n \sim P_\theta$, for some function c such that $0 < c(\theta_A) < \infty$ for $\theta_A \in \Theta \backslash \Theta_0$, and $c(\theta_0) = 0$ when $\theta_0 \in \Theta_0$. The function $c(\theta)$ is known as the *approximate Bahadur slope* (ABS) of the sequence T_n. The quantifier "approximate" comes from the use of the asymptotic null distribution instead of the exact one [Bahadur, 1960]. Intuitively the slope $c(\theta_A)$, for $\theta_A \in \Theta \backslash \Theta_0$, is the rate of convergence of p-values (i.e., $1 - F(T_n)$) to 0, as n increases. The higher the

slope, the faster the p-value vanishes, and thus the lower the sample size required to reject H_0 under θ_A.

Approximate Bahadur Efficiency Given two sequences of test statistics, $T_n^{(1)}$ and $T_n^{(2)}$ having the same $\rho(n)$ (see Theorem 5.8), the approximate Bahadur efficiency of $T_n^{(1)}$ relative to $T_n^{(2)}$ is defined as $E(\theta_A) := c^{(1)}(\theta_A)/c^{(2)}(\theta_A)$ for $\theta_A \in \Theta \backslash \Theta_0$. If $E(\theta_A) > 1$, then $T_n^{(1)}$ is asymptotically more efficient than $T_n^{(2)}$ in the sense of Bahadur, for the particular problem specified by $\theta_A \in \Theta \backslash \Theta_0$.

In practice, the main difficulty in determining the approximate Bahadur slope is the computation of $-2 \operatorname{plim}_{n\to\infty} \frac{\log(1-F(T_n))}{\rho(n)}$, typically requiring the aid of the theory of large deviations. There are further sufficient conditions which make the computation easier. The following conditions are due to Gleser [1964, 1966], first appearing in Bahadur [1960] in a slightly less general form.

Definition 5.6. Let $\mathcal{D}(a,t)$ be a class of all continuous cumulative distribution functions (CDF) F such that $-2\log(1-F(x)) = ax^t(1+o(1))$, as $x \to \infty$ for $a > 0$ and $t > 0$.

Theorem 5.7 (Gleser [1964, 1966])**.** *Consider a sequence of test statistic T_n. Assume that*

1. *There exists a function $F(x)$ such that for $\theta \in \Theta_0$, $\lim_{n\to\infty} P_\theta(T_n < x) = F(x)$, for all x, and such that $F \in \mathcal{D}(a,t)$ for some $a > 0$ and $t > 0$ (see Definition 5.6).*
2. *There exists a continuous, strictly increasing function $R : (0,\infty) \to (0,\infty)$ with $\lim_{n\to\infty} R(n) = \infty$, and a function $b(\theta)$ with $0 < b(\theta) < \infty$ defined on $\Theta \backslash \Theta_0$, such that for all $\theta \in \Theta \backslash \Theta_0$, $\operatorname{plim}_{n\to\infty} T_n/R(n) = b(\theta)$.*

Then, $-2 \operatorname{plim}_{n\to\infty} \frac{\log(1-F(T_n))}{[R(n)]^t} = a\,[b(\theta)]^t =: c(\theta)$, the approximate slope of the sequence T_n, where $\rho(n) = R(n)^t$ (see Section 5.4).

Theorem 5.8 (Gleser [1964, 1966])**.** *Consider two sequences of test statistics $T_n^{(1)}$ and $T_n^{(2)}$. Let $F^{(i)}$ be the CDF of $T_n^{(i)}$ for $i = 1, 2$. Assume that each sequence satisfies all the conditions in Theorem 5.7 with $F^{(i)} \in \mathcal{D}(a_i, t_i)$. Further, assume that $\left[R^{(1)}(x)\right]^{t_1} = \left[R^{(2)}(x)\right]^{t_2}$ for all x. Then*

$$\operatorname*{plim}_{n\to\infty} \frac{\log(1-F^{(1)}(T_n^{(1)}))}{\log(1-F^{(2)}(T_n^{(2)}))} = \frac{c^{(1)}(\theta)}{c^{(2)}(\theta)} = \varphi_{1,2}(\theta),$$

which is the approximate Bahadur efficiency of $T_n^{(1)}$ relative to $T_n^{(2)}$.

With Theorem 5.7, the difficulty is in showing that $F \in \mathcal{D}(a,t)$ for some $a > 0, t > 0$. Typically verification of the assumption 2 of Theorem 5.7 poses no problem. Bahadur [1960] showed that the CDF of $\mathcal{N}(0,1)$ belongs to $\mathcal{D}(1,2)$ and the CDF of χ_k^2 (chi-squared distribution with k degrees of freedom, fixed k) belongs to $\mathcal{D}(1,1)$. The following theorem makes it easier to determine whether a given CDF is in the class $\mathcal{D}(a,t)$.

Theorem 5.9 (Gleser [1966, Theorem 6, 7]). *Let X have CDF $F \in \mathcal{D}(a,t)$, and $X_1,\dots,X_m$ be independent random variables, each with CDF $F_i \in \mathcal{D}(a,t)$. Then, the following statements are true.*

1. *If $b > 0$, then the CDF of bX is in $\mathcal{D}(ab^{-t},t)$.*
2. *$X - b$ has CDF in $\mathcal{D}(a,t)$ provided that $t \geq 1$.*
3. *For $r > 0$, X^r has CDF in $\mathcal{D}(a,r^{-1}t)$ provided that $F(0) = 0$.*
4. *$\max(X_1,\dots,X_m)$ has CDF in $\mathcal{D}(a,t)$.*
5. *Let $a_1,\dots,a_m$ be non-negative real numbers such that $a_{max} := \max(a_1,\dots,a_m) > 0$. Then, $\sum_{i=1}^m a_i X_i$ has CDF in $\mathcal{D}(a \cdot a_{max}^{-t}, t)$ provided that $\sum_{i=1}^m X_i$ has CDF in $\mathcal{D}(a,t)$ and $X_i \geq 0$ for all $i = 1,\dots,m$.*

5.4.2 Approximate Bahadur Slopes of $n\widehat{\text{FSSD}^2}$ and $\sqrt{n}\widehat{S}_l^2$

We now give approximate Bahadur slopes for two sequences of linear time test statistics: the proposed $n\widehat{\text{FSSD}^2}$ (Theorem 5.10), and the LKS test statistic $\sqrt{n}\widehat{S}_l^2$ (Theorem 5.11). We then show in Theorem 5.14 that when $p = \mathcal{N}(0,1)$ and $q = \mathcal{N}(\mu_q, 1)$, the approximate Bahadur efficiency of $n\widehat{\text{FSSD}^2}$ relative to $\sqrt{n}\widehat{S}_l^2$ is always greater than 2 for appropriately chosen hyperparameters (i.e., a Gaussian kernel bandwidth and a test location).

Theorem 5.10. *The approximate Bahadur slope of $n\widehat{\text{FSSD}^2}$ is $c^{(\text{FSSD})} := \text{FSSD}^2/\omega_1$, where ω_1 is the maximum eigenvalue of $\boldsymbol{\Sigma}_p := \mathbb{E}_{\mathbf{x}\sim p}[\boldsymbol{\tau}(\mathbf{x})\boldsymbol{\tau}^\top(\mathbf{x})]$ (see Section 5.3.1 for the definition of $\boldsymbol{\tau}$) and $\rho(n) = n$ (see Section 5.4.1 for ρ).*

Proof. We will use Theorem 5.7 to derive the slope. For the assumption 1 of Theorem 5.7, we first show that the asymptotic null distribution belongs to the class $\mathcal{D}(a = 1/\omega_1, t = 1)$ as defined in Definition 5.6. By Proposition 5.3, the asymptotic null distribution is $\sum_{i=1}^{dJ} \omega_i Z_i^2 - \sum_{i=1}^{dJ} \omega_i$ where $Z_1,\dots,Z_{dJ} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $\omega_1 \geq \cdots \geq \omega_{dJ} \geq 0$ are eigenvalues of $\boldsymbol{\Sigma}_p$. It is known from Bahadur [1960] that the CDF of χ_f^2 is in $\mathcal{D}(1,1)$ for any fixed degrees of freedom f. Thus, it follows from claim 5 of Theorem 5.9 that the CDF of $\sum_{i=1}^{dJ} \omega_i Z_i^2$ is in $\mathcal{D}(a = 1/\omega_1, t = 1)$. Claim 2 of Theorem 5.9 guarantees that the CDF of $\sum_{i=1}^{dJ} \omega_i Z_i^2 - \sum_{i=1}^{dJ} \omega_i$ is in $\mathcal{D}(a = 1/\omega_1, t = 1)$ as desired.

For assumption 2 of Theorem 5.7, choose $R(n) := n$. It follows from the weak law of large numbers that under H_1, $n\widehat{\text{FSSD}^2}/R(n) \xrightarrow{p} \text{FSSD}^2$. By Theorem 5.7, the approximate slope is FSSD^2/ω_1. □

Theorem 5.11. *The approximate Bahadur slope of the linear-time kernel Stein (LKS) test statistic $\sqrt{n}\widehat{S}_l^2$ is $c^{(\text{LKS})} = \frac{1}{2}\frac{\left[\mathbb{E}_q h_p(\mathbf{x},\mathbf{x}')\right]^2}{\mathbb{E}_p\left[h_p^2(\mathbf{x},\mathbf{x}')\right]}$, where h_p is the U-statistic kernel in (5.2) of the KSD statistic, and $\rho(n) = n$.*

Proof. We will use Theorem 5.7 to derive the slope. By the central limit theorem,

$$\sqrt{n}\left(\widehat{S_l^2} - S_p^2(q)\right) \xrightarrow{d} \mathcal{N}(0, 2\mathbb{V}_q[h_p(\mathbf{x}, \mathbf{x}')]),$$

where $\mathbb{V}_q[h_p(\mathbf{x}, \mathbf{x}')] := \mathbb{E}_{\mathbf{x}\sim q}\mathbb{E}_{\mathbf{x}'\sim q}[h_p^2(\mathbf{x}, \mathbf{x}')] - \left(\mathbb{E}_{\mathbf{x}\sim q}\mathbb{E}_{\mathbf{x}'\sim q}[h_p(\mathbf{x}, \mathbf{x}')]\right)^2$. Under $H_0 : p = q$, it follows that $S_p^2(q) = \mathbb{E}_{\mathbf{x}\sim q}\mathbb{E}_{\mathbf{x}'\sim q}[h_p(\mathbf{x}, \mathbf{x}')] = 0$ by Theorem 5.1, and $\sqrt{n}\widehat{S_l^2} \xrightarrow{d} \mathcal{N}(0, 2\mathbb{V}_p[h_p(\mathbf{x}, \mathbf{x}')])$ where $\mathbb{V}_p[h_p(\mathbf{x}, \mathbf{x}')] := \mathbb{E}_{\mathbf{x}\sim p}\mathbb{E}_{\mathbf{x}'\sim p}[h_p^2(\mathbf{x}, \mathbf{x}')]$. It is known from Bahadur [1960] that the CDF of $\mathcal{N}(0, 1)$ is in the class $\mathcal{D}(1, 2)$ (see Definition 5.6). Thus, by property 1 of Theorem 5.9, the CDF of $\mathcal{N}(0, 2\mathbb{V}_p[h_p(\mathbf{x}, \mathbf{x}')])$ is in $\mathcal{D}\left(a = \frac{1}{2\mathbb{V}_p[h_p(\mathbf{x},\mathbf{x}')]}, t = 2\right)$.

For assumption 2 of Theorem 5.7, choose $R(n) := \sqrt{n}$. It follows from the weak law of large numbers that under H_1, $\sqrt{n}\widehat{S_l^2}/R(n) = \widehat{S_l^2} \xrightarrow{p} S_p^2(q)$. By Theorem 5.7, the approximate slope is $\frac{S_p^4(q)}{2\mathbb{V}_p[h_p(\mathbf{x},\mathbf{x}')]}$ implying the result. □

To make these results concrete, we consider the setting where $p = \mathcal{N}(0, 1)$ and $q = \mathcal{N}(\mu_q, \sigma_q^2)$ on the real line. We assume that both tests use the Gaussian kernel $k(x, y) = \exp\left(-(x - y)^2/2\sigma_k^2\right)$, possibly with different bandwidths. We write σ_k^2 and κ^2 for the kernel bandwidths of FSSD and LKS, respectively. Under these assumptions, the slopes given in Theorem 5.10 and Theorem 5.11 can be derived explicitly. The full expressions of the slopes are given in Proposition 5.12 and Proposition 5.13.

Proposition 5.12. *Under the setting that $J = 1$ (i.e., one test location v), $p = \mathcal{N}(0, 1)$ and $q = \mathcal{N}(\mu_q, \sigma_q^2)$, the approximate Bahadur Slope of $n\widehat{\text{FSSD}^2}$ is*

$$c^{(\text{FSSD})} := \frac{\left(\sigma_k^2\right)^{3/2}\left(\sigma_k^2 + 2\right)^{5/2} e^{\frac{v^2}{\sigma_k^2+2} - \frac{(v-\mu_q)^2}{\sigma_k^2+\sigma_q^2}}\left(\left(\sigma_k^2 + 1\right)\mu_q + v\left(\sigma_q^2 - 1\right)\right)^2}{\left(\sigma_k^2 + \sigma_q^2\right)^3\left(\sigma_k^6 + 4\sigma_k^4 + \left(v^2 + 5\right)\sigma_k^2 + 2\right)}. \tag{5.6}$$

Proof. This result follows directly from Theorem 5.10 specialized to the case of $p = \mathcal{N}(0, 1)$, $q = \mathcal{N}(\mu_q, \sigma_q^2)$, and $J = 1$. Since $dJ = 1$, the covariance matrix

$$\mathbf{\Sigma}_p = \mathbb{E}_{x\sim p}\left[\xi_p^2(x, v)\right] = \frac{e^{-\frac{v^2}{\sigma_k^2+2}}\left(\sigma_k^6 + 4\sigma_k^4 + \left(v^2 + 5\right)\sigma_k^2 + 2\right)}{\sigma_k\left(\sigma_k^2 + 2\right)^{5/2}}$$

reduces to a scalar, where

$$\begin{aligned}\xi_p(x, v) &= \left[\frac{\partial}{\partial x}\log p(x)\right]k(x, v) + \frac{\partial}{\partial x}k(x, v) \\ &= -e^{-\frac{(v-x)^2}{2\sigma_k^2}}\left(x\sigma_k^2 - v + x\right)/\sigma_k^2.\end{aligned}$$

In this case,

$$\text{FSSD}^2 = \mathbb{E}_{x\sim q}^2\left[\xi_p(x, v)\right] = \frac{\sigma_k^2 e^{-\frac{(v-\mu_q)^2}{\sigma_k^2+\sigma_q^2}}\left(\left(\sigma_k^2 + 1\right)\mu_q + v\left(\sigma_q^2 - 1\right)\right)^2}{\left(\sigma_k^2 + \sigma_q^2\right)^3}. \tag{5.7}$$

Taking the ratio $\mathrm{FSSD}^2 / \mathbb{E}_{x\sim p}\left[\xi_p^2(x,v)\right]$ gives the result. □

Remark 2. As noted in Theorem 5.2 that $\mathrm{FSSD}_p^2(q) = 0$ if and only if $p = q$ almost surely with respect to the distribution η with a density, from which v is drawn. The η-almost sureness means that when $p \neq q$,

1. There might exist a set V_0 of measure zero such that $\mathrm{FSSD}_p^2(q) = 0$ when $v \in V_0$.
2. When $v \sim \eta$, $v \notin V_0$ with probability 1 and hence $\mathrm{FSSD}_p^2(q) > 0$.

To concretely demonstrate this, consider the case where $p = \mathcal{N}(0,1)$, $q = \mathcal{N}(\mu_q, \sigma_q^2)$, $\mu_q \neq 0$ and $\sigma_q^2 \neq 1$. From (5.7), it can be seen that $V_0 = \left\{-\left(\sigma_k^2+1\right)\mu_q / \left(\sigma_q^2-1\right)\right\}$ containing only one element. Since η has a density, the probability that v drawn from η is realized to be this value is 0.

Proposition 5.13. *Assume that $p = \mathcal{N}(0,1)$ and $q = \mathcal{N}(\mu_q, \sigma_q^2)$. Let $\sqrt{n}\widehat{S}_l^2$ be the linear-time kernel Stein (LKS) test statistic where $\widehat{S}_l^2$ is defined in (5.3) with a Gaussian kernel $k(x,y) = \exp\left(-\frac{(x-y)^2}{2\kappa^2}\right)$. Then, the following statements hold.*

1. *The population kernel Stein discrepancy is*

$$S_p^2(q) = \frac{\mu_q^2\left(\kappa^2 + 2\sigma_q^2\right) + \left(\sigma_q^2 - 1\right)^2}{\left(\kappa^2 + 2\sigma_q^2\right)\sqrt{\frac{2\sigma_q^2}{\kappa^2}+1}}.$$

2. *The approximate Bahadur slope of $\sqrt{n}\widehat{S}_l^2$ is*

$$c^{(\mathrm{LKS})} := \frac{\kappa^5\left(\kappa^2+4\right)^{5/2}\left[\mu_q^2\left(\kappa^2+2\sigma_q^2\right) + \left(\sigma_q^2-1\right)^2\right]^2}{2\left(\kappa^8+8\kappa^6+21\kappa^4+20\kappa^2+12\right)\left(\kappa^2+2\sigma_q^2\right)^3}. \tag{5.8}$$

3. *Let*

$$c_1^{(\mathrm{LKS})} = \frac{\left(\kappa^2\right)^{5/2}\left(\kappa^2+4\right)^{5/2}\mu_q^4}{2\left(\kappa^2+2\right)\left(\kappa^8+8\kappa^6+21\kappa^4+20\kappa^2+12\right)}$$

denote the approximate slope $c^{(\mathrm{LKS})}$ specialized to when $q = \mathcal{N}(\mu_q, 1)$. Then, for any $\mu_q \neq 0$, the function $\kappa^2 \mapsto c_1^{(\mathrm{LKS})}(\mu_q, \kappa^2)$ is strictly increasing on $(0,\infty)$. Further,

$$\lim_{\kappa^2\to\infty} c_1^{(\mathrm{LKS})}(\mu_q, \kappa^2) = \mu_q^4/2. \tag{5.9}$$

Proof. **Proof of Claim 1, 2.** Recall $\widehat{S}_l^2 := \frac{2}{n}\sum_{i=1}^{n/2} h_p(x_{2i-1}, x_{2i})$. With $p = \mathcal{N}(0,1)$, and $k(x,y) = \exp\left(-\frac{(x-y)^2}{2\kappa^2}\right)$, $h_p(x,y)$ can be written as

$$h_p(x,y) := \frac{e^{-\frac{(x-y)^2}{2\kappa^2}}\left(\kappa^2 - \left(\kappa^2+1\right)x^2 + \left(\kappa^4+2\kappa^2+2\right)xy - \left(\kappa^2+1\right)y^2\right)}{\kappa^4}.$$

By Theorem 5.11, $c^{(\mathrm{LKS})} = \frac{1}{2}\frac{\left[\mathbb{E}_q h_p(\mathbf{x},\mathbf{x}')\right]^2}{\mathbb{E}_p\left[h_p^2(\mathbf{x},\mathbf{x}')\right]}$ which mainly involves expectations with respect to a normal distribution. In computing the expectation $\mathbb{E}_{x'\sim q} h_p(x,x')$, the idea is to form the density for a new normal distribution by combining $\frac{1}{\sqrt{2\pi\sigma_q^2}} e^{-(x-\mu_q)^2/2\sigma_q^2}$ (the density of q) and the term $e^{-\frac{(x-y)^2}{2\kappa^2}}$ in the expression of $h_p(x,y)$. Computation of $\mathbb{E}_{x'\sim q} h_p(x,x')$ will then boil down to computing an expectation wrt. a new normal distribution.

It turns out that

$$\mathbb{E}_{x\sim q}\mathbb{E}_{x'\sim q}[h_p(x,x')] = \frac{\mu_q^2\left(\kappa^2+2\sigma_q^2\right)+\left(\sigma_q^2-1\right)^2}{\left(\kappa^2+2\sigma_q^2\right)\sqrt{\frac{2\sigma_q^2}{\kappa^2}+1}} = S_p^2(q),$$
$$\mathbb{E}_p\left[h_p^2(\mathbf{x},\mathbf{x}')\right] = \frac{\left(\kappa^2+4\right)\left(\kappa^4+4\kappa^2+5\right)\kappa^2+12}{\kappa^3\left(\kappa^2+4\right)^{5/2}}.$$

Computing $\frac{1}{2}\frac{S_p^4(q)}{\mathbb{E}_p\left[h_p^2(\mathbf{x},\mathbf{x}')\right]}$ gives the slope.

Proof of Claim 3. The expression for $c_1^{(\mathrm{LKS})}$ is obtained straightforwardly by plugging $\sigma_q^2 = 1$ into the expression of $c^{(\mathrm{LKS})}$. Assume $\mu_q \neq 0$. It can be seen that $c_1^{(\mathrm{LKS})}(\mu_q,\kappa^2)$ is differentiable with respect to κ^2 on the interval $(0,\infty)$. The partial derivative is given by

$$\frac{\partial}{\partial\kappa^2}c_1^{(\mathrm{LKS})} = \frac{\left(\kappa^2\right)^{3/2}\left(\kappa^2+4\right)^{3/2}\left(7\kappa^8+56\kappa^6+166\kappa^4+216\kappa^2+120\right)\mu_q^4}{\left(\kappa^2+2\right)^2\left(\kappa^8+8\kappa^6+21\kappa^4+20\kappa^2+12\right)^2}.$$

Since for any $\mu_q \neq 0$, $\frac{\partial}{\partial\kappa^2}c_1^{(\mathrm{LKS})} > 0$ for $\kappa^2 \in (0,\infty)$, we conclude that $\kappa^2 \mapsto c_1^{(\mathrm{LKS})}(\mu_q,\kappa^2)$ is a strictly increasing function on $(0,\infty)$. By taking the limit, we have $\lim_{\kappa^2\to\infty} c_1^{(\mathrm{LKS})}(\mu_q,\kappa^2) = \mu_q^4/2$. □

By Theorem 5.8, the approximate Bahadur efficiency can be computed by taking the ratio of the two slopes. For this purpose, we consider $p = \mathcal{N}(0,1)$ and $q = \mathcal{N}(\mu_q,1)$ i.e., a mean shift problem. The efficiency is given in Theorem 5.14.

Theorem 5.14 (Bahadur efficiency in the Gaussian mean shift problem)**.** *Let $E_1(\mu_q,v,\sigma_k^2,\kappa^2)$ be the approximate Bahadur efficiency of $n\widehat{\mathrm{FSSD}^2}$ relative to $\sqrt{n}\widehat{S}_l^2$ for the case where $p = \mathcal{N}(0,1), q = \mathcal{N}(\mu_q,1)$, and $J = 1$ (i.e., one test location v for $n\widehat{\mathrm{FSSD}^2}$). Fix $\sigma_k^2 = 1$ for $n\widehat{\mathrm{FSSD}^2}$. Then, for any $\mu_q \neq 0$, for some $v \in \mathbb{R}$, and for any $\kappa^2 > 0$, we have $E_1(\mu_q,v,\sigma_k^2,\kappa^2) > 2$.*

Proof. By Proposition 5.12, the approximate slope of $n\widehat{\mathrm{FSSD}^2}$ when $\sigma_q^2 = 1$ is

$$c_1^{(\mathrm{FSSD})}(\mu_q,v,\sigma_k^2) = \frac{\sigma_k^2\left(\sigma_k^2+2\right)^3\mu_q^2 e^{\frac{v^2}{\sigma_k^2+2}-\frac{(v-\mu_q)^2}{\sigma_k^2+1}}}{\sqrt{\frac{2}{\sigma_k^2}+1}\left(\sigma_k^2+1\right)\left(\sigma_k^6+4\sigma_k^4+\left(v^2+5\right)\sigma_k^2+2\right)}.$$

Theorem 5.8 states that the approximate efficiency $E_1(\mu_q, v, \sigma_k^2, \kappa^2)$ is given by the ratio $\frac{c_1^{(\mathrm{FSSD})}(\mu_q, v, \sigma_k^2)}{c_1^{(\mathrm{LKS})}(\mu_q, \kappa^2)}$ (see Propositions 5.12 and 5.13) of the approximate slopes of the two tests. Pick $\sigma_k^2 = 1$, and for any $\mu_q \neq 0$, pick $v = 2\mu_q$. These choices give the slope

$$c_1^{(\mathrm{FSSD})}(\mu_q, 2\mu_q, 1) = \frac{9\sqrt{3}e^{\frac{5\mu_q^2}{6}}\mu_q^2}{2\left(4\mu_q^2 + 12\right)}.$$

We have

$$\begin{aligned} E_1(\mu_q, v, \sigma_k^2, \kappa^2) &= E_1(\mu_q, 2\mu_q, 1, \kappa^2) \\ &= c_1^{(\mathrm{FSSD})}(\mu_q, 2\mu_q, 1) / c_1^{(\mathrm{LKS})}(\mu_q, \kappa^2) \\ &\overset{(a)}{\geq} c_1^{(\mathrm{FSSD})}(\mu_q, 2\mu_q, 1) / \left(\frac{\mu_q^4}{2}\right) \\ &= \frac{9\sqrt{3}e^{\frac{5\mu_q^2}{6}}}{\mu_q^2\left(4\mu_q^2 + 12\right)} := g(\mu_q), \end{aligned}$$

where at (a) we use $c_1^{(\mathrm{LKS})}(\mu_q, \kappa^2) \leq \mu_q^4/2$ from (5.9). It can be seen that for $\mu_q \neq 0$, $g(\mu_q)$ is an even function i.e., $g(\mu_q) = g(-\mu_q)$. The second derivative

$$\frac{\partial^2}{\partial \mu_q^2} g(\mu_q) = \sqrt{3}e^{\frac{5\mu_q^2}{6}}\left(25\mu_q^8 + 45\mu_q^6 - 45\mu_q^4 + 81\mu_q^2 + 486\right) / \left(4\mu_q^4\left(\mu_q^2 + 3\right)^3\right) > 0.$$

To see that $\frac{\partial^2}{\partial \mu_q^2} g(\mu_q) > 0$, consider two cases of $\mu_q^2 \geq 1$ and $0 < \mu_q^2 < 1$. When $\mu_q^2 \geq 1$,

$$g(\mu_q) \geq \sqrt{3}e^{\frac{5\mu_q^2}{6}}\left(25\mu_q^8 + 81\mu_q^2 + 486\right) / \left(4\mu_q^4\left(\mu_q^2 + 3\right)^3\right) > 0,$$

because $45\mu_q^6 - 45\mu_q^4 \geq 0$. When $0 < \mu_q^2 < 1$,

$$g(\mu_q) \geq \sqrt{3}e^{\frac{5\mu_q^2}{6}}\left(25\mu_q^8 + 45\mu_q^6 + 486\right) / \left(4\mu_q^4\left(\mu_q^2 + 3\right)^3\right) > 0,$$

because $-45\mu_q^4 + 81\mu_q^2 \geq 0$. This shows that $g(\mu_q)$ is convex on $(0, \infty)$. The function $g(\mu_q)$ on $\mathbb{R}\backslash\{0\}$ achieves global minima at $\mu_q = \mu_q^* := \pm\sqrt{\frac{3}{10}\left(\sqrt{41} - 1\right)} \approx \pm 1.273$. This implies that

$$\begin{aligned} E_1(\mu_q, v, \sigma_k^2, \kappa^2) &\geq g(\mu_q) \geq g(\mu_q^*) \\ &= \frac{25\sqrt{3}e^{\frac{1}{4}\left(\sqrt{41}-1\right)}}{8\left(\sqrt{41} + 4\right)} \approx 2.00855 > 2. \end{aligned}$$

□

When $p = \mathcal{N}(0,1)$ and $q = \mathcal{N}(\mu_q, 1)$ for $\mu_q \neq 0$, Theorem 5.14 guarantees that our FSSD test is asymptotically at least twice as efficient as the LKS test in the Bahadur sense. We note that the efficiency is conservative in the sense that $\sigma_k^2 = 1$ regardless of μ_q. Choosing σ_k^2 dependent on μ_q will likely improve the efficiency further.

5.5 Experiments

In this section, we demonstrate the performance of the proposed test on a number of problems. The primary goal is to understand the conditions under which the test can perform well. Our goal is *not* to demonstrate that the proposed test outperforms the quadratic-time test of Chwialkowski et al. [2016], Liu et al. [2016] over all possible problems.

5.5.1 Sensitivity to Local Differences

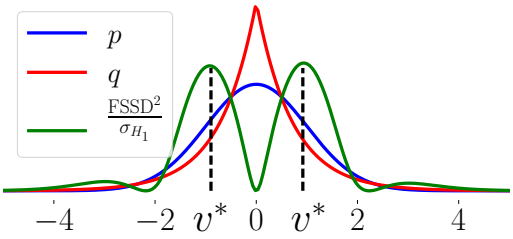


Figure 5.1: The power criterion $\text{FSSD}^2/\sigma_{H_1}$ as a function of test location v, when $p = \mathcal{N}(0,1)$ and $q = \text{Laplace}(0, 1/\sqrt{2})$.

We start by demonstrating that the test power objective $\text{FSSD}^2/\sigma_{H_1}$ captures local differences of p and q, and that interpretable features v are found. Consider a one-dimensional problem in which $p = \mathcal{N}(0,1)$ and $q = \text{Laplace}(0, 1/\sqrt{2})$, a zero-mean Laplace distribution with scale parameter $1/\sqrt{2}$. These parameters are chosen so that p and q have the same mean and variance. Figure 5.1 plots the (rescaled) objective as a function of v. The objective illustrates that the best features (indicated by v^*) are at the most discriminative locations.

5.5.2 Test Power

We next investigate the power of different tests on two problems with different characteristics:

1. **Gaussian vs. Laplace**: $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{I}_d)$ and $q(\mathbf{x}) = \prod_{i=1}^{d} \text{Laplace}(x_i|0, 1/\sqrt{2})$ where the dimension d will be varied. The two distributions have the same mean and variance. The main characteristic of this problem is local differences of p and q (see Figure 5.1). Set $n = 1000$.
2. **Restricted Boltzmann Machine** (RBM): $p(\mathbf{x})$ is the marginal distribution of

$$p(\mathbf{x}, \mathbf{h}) = \frac{1}{Z} \exp\left(\mathbf{x}^\top \mathbf{B}\mathbf{h} + \mathbf{b}^\top \mathbf{x} + \mathbf{c}^\top \mathbf{x} - \frac{1}{2}\|\mathbf{x}\|^2\right),$$

where $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{h} \in \{\pm 1\}^{d_h}$ is a random vector of hidden variables, and Z is the normalization constant. The exact marginal density $p(\mathbf{x}) = \sum_{\mathbf{h} \in \{-1,1\}^{d_h}} p(\mathbf{x}, \mathbf{h})$ is intractable when d_h is large, since it involves summing over 2^{d_h} terms. Recall that the proposed test only requires the score function $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ (not the normalization constant), which can be computed in closed form in this case, and is given by

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}) = \mathbf{b} - \mathbf{x} + \mathbf{B}\varphi(\mathbf{B}^\top \mathbf{x} + \mathbf{c}),$$
$$\varphi(z) = \frac{\exp(2z) - 1}{\exp(2z) + 1},$$

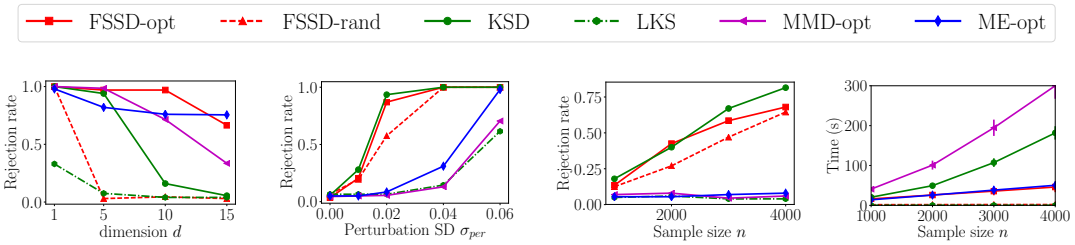
where φ applies element-wise to a vector input [Liu et al., 2016, Section 6]. In this problem, q is another RBM where entries of the matrix $\mathbf{B}$ are corrupted by Gaussian noise. This problem was considered in Liu et al. [2016]. We set $d = 50$ and $d_h = 40$, and generate samples by n independent chains (i.e., n independent samples) of blocked Gibbs sampling with 2000 burn-in iterations.

We evaluate the following six kernel-based nonparametric tests with $\alpha = 0.05$, all using the Gaussian kernel.

1. **FSSD-rand**: the proposed FSSD test where the test locations set to random draws from a multivariate normal distribution fitted to the data. The kernel bandwidth is set by the commonly used median heuristic i.e., $\sigma_k = \text{median}(\{\|\mathbf{x}_i - \mathbf{x}_j\|, i < j\})$.
2. **FSSD-opt**: the proposed FSSD test where both the test locations and the Gaussian bandwidth are optimized (Section 5.3.2). For both the FSSD tests, we use $J = 5$.
3. **KSD**: the quadratic-time Kernel Stein Discrepancy test with the median heuristic.
4. **LKS**: the linear-time version of KSD with the median heuristic.
5. **MMD-opt**: the quadratic-time MMD two-sample test of Gretton et al. [2012a] where the kernel bandwidth is optimized by grid search to maximize a power criterion as described in Sutherland et al. [2016].
6. **ME-opt**: the linear-time mean embeddings (ME) two-sample test of Jitkrittum et al. [2016] where parameters are optimized. We draw n samples from p to run the two-sample tests (MMD-opt, ME-opt).

All tests with optimization use 20% of the sample size n for parameter tuning. Each problem is repeated for 200 trials, resampling n points from q every time.

Gaussian vs. Laplace In Figure 5.2a (Gaussian vs. Laplace), high performance of FSSD-opt indicates that the test performs well when there are local differences between p and q. Low performance of FSSD-rand emphasizes the importance of the optimization of FSSD-opt to pinpoint regions where p and q differ. The power of KSD



(a) Gaussian vs. Laplace. $n = 1000$.

(b) RBM. $n = 1000$. Perturb all entries of $\mathbf{B}$.

(c) RBM. $\sigma_{per} = 0.1$. Perturb $B_{1,1}$.

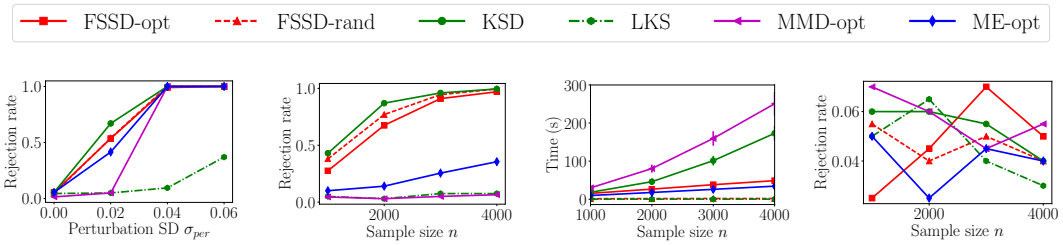
(d) Runtime (RBM)

Figure 5.2: Rejection rates of the six tests in the two problems. In the RBM problem, $d = 50$ and $d_h = 40$. The proposed linear-time FSSD-opt has a comparable or higher test power in some cases than the quadratic-time KSD test.

quickly drops as the dimension increases, which can be understood since KSD is the RKHS norm of a function witnessing differences in p and q across the entire domain, including where these differences are small.

RBM with $d = 50$ and $d_h = 40$ We next consider the case of RBMs with $d = 50$ and $d_h = 40$. Following Liu et al. [2016], $\mathbf{b}, \mathbf{c}$ are independently drawn from the standard multivariate normal distribution, and entries of $\mathbf{B} \in \mathbb{R}^{50\times 40}$ are drawn with equal probability from $\{\pm 1\}$, in each trial. The density q represents another RBM having the same $\mathbf{b}, \mathbf{c}$ as in p, and with all entries of $\mathbf{B}$ corrupted by independent zero-mean Gaussian noise with standard deviation σ_{per}. Figure 5.2b shows the test powers as σ_{per} increases, for a fixed sample size $n = 1000$. We observe that all the tests have correct false positive rates (type-I errors) at roughly $\alpha = 0.05$ when there is no perturbation noise. In particular, the optimization in FSSD-opt does not increase false positive rate when H_0 holds. We see that the performance of the proposed FSSD-opt matches that of the quadratic-time KSD at all noise levels. MMD-opt and ME-opt perform far worse than the goodness-of-fit tests when the difference in p and q is small (σ_{per} is low), since these tests simply represent p using samples, and do not take advantage of its structure.

The advantage of having $\mathcal{O}(n)$ runtime can be clearly seen when the problem is much harder, requiring larger sample sizes to tackle. Consider a similar problem on RBMs in which the parameter $\mathbf{B} \in \mathbb{R}^{50\times 40}$ in q is given by that of p, where only the first entry $B_{1,1}$ is perturbed by random $\mathcal{N}(0, 0.1^2)$ noise. The results are shown in Figure 5.2c where the sample size n is varied. We observe that the two two-sample tests fail to detect this subtle difference even with large sample size. The test powers of KSD and FSSD-opt are comparable when n is relatively small. It appears that KSD has higher test power than FSSD-opt in this case for large n. However, this moderate gain in the test power comes with an order of magnitude more computation. As shown in Figure 5.2d, the runtime of the KSD is much larger than that of FSSD-opt, especially at large n. In these problems, the performance of the new test (even without optimization) far exceeds that of the LKS test.



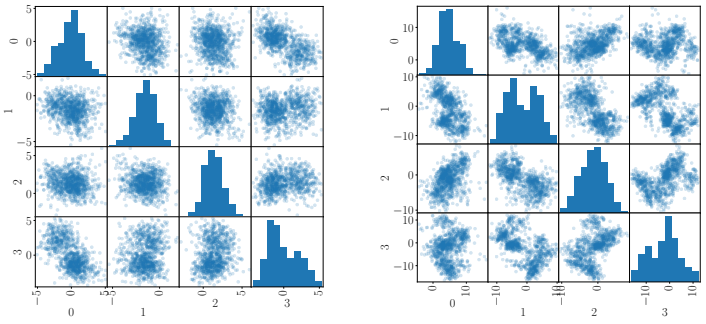
(a) RBM. $n = 1000$. Perturb all entries of $\mathbf{B}$.

(b) RBM. $\sigma_{per} = 0.1$. Perturb $B_{1,1}$.

(c) Runtime (RBM)

(d) RBM. No perturbation. H_0 holds.

Figure 5.3: Rejection rates of the six tests in the RBM problem with $d = 50$ and $d_h = 10$.



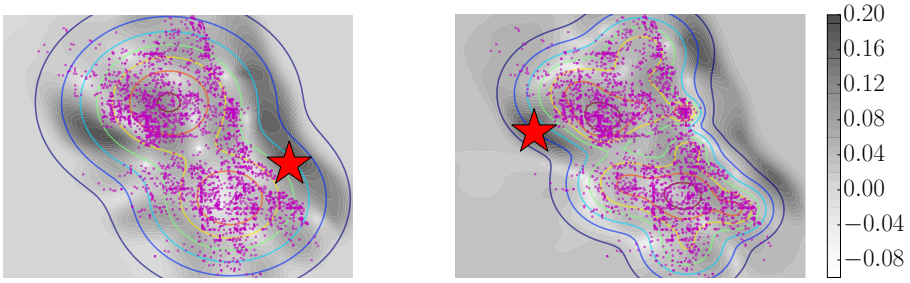
(a) $d = 50, d_h = 10$

(b) $d = 50, d_h = 40$

Figure 5.4: Pairwise scatter plots of 1000 points drawn from RBMs. Only the first 4 variates out of 50 are shown. **(a)**: RBM with $d = 50$ dimensions with $d_h = 10$ latent variables. **(b)**: RBM with $d = 50$ dimensions with $d_h = 40$ latent variables.

RBM with $d = 50$ and $d_h = 10$ In Liu et al. [2016], the setting of $d = 50$ and $d_h = 10$ was studied. For completeness, we consider the same setting and show the results in Figure 5.3 where all other problem configurations are the same.

In Figure 5.3a, p is set to an RBM with parameters randomly drawn (described in Section 5.5), and q is the same RBM with all entries of the parameter $\mathbf{B} \in \mathbb{R}^{50\times 10}$ perturbed by independent Gaussian noise with standard deviation σ_{per}, which varies from 0 to 0.06. We observe that the proposed FSSD-opt and KSD perform comparably. Figure 5.3b considers a hard problem where only the first entry $B_{1,1}$ is perturbed by noise following $\mathcal{N}(0, 0.1^2)$, and the sample size n is varied. In both of these two cases, the overall trend is similar to the case of $d = 50$ and $d_h = 40$ presented in Figure 5.2. It is interesting to note that FSSD-rand, relying on random test locations, performs comparably or even outperforms FSSD-opt in the case of $d = 50, d_h = 10$, but not in the case of $d = 50, d_h = 40$. This phenomenon can be explained as follows. In the case of $d = 50, d_h = 10$, the data generated from the RBM tend to have simple structure (see Figure 5.4a). By contrast, data generated from the RBM with $d = 50, d_h = 40$ (more latent variables) have larger variance, and can form a complicated structure (Figure 5.4b), requiring a careful choice of test locations to detect differences of p and q. When $d = 50, d_h = 10$, however, random test locations given by random draws from



(a) $p = 2$-component GMM. (b) $p = 10$-component GMM

Figure 5.5: Plots of the optimization objective as a function of test location $\mathbf{v} \in \mathbb{R}^2$ in the Gaussian mixture model (GMM) evaluation task. Density functions of the fitted GMMs are shown in wireframe. The parameter tuning objective is shown as a grayscale contour plot.

a Gaussian distribution fitted to the data are sufficient to capture the simple structural difference. This explains why FSSD-rand can perform well in this case. Additionally, FSSD-rand also has 20% more testing data, since FSSD-opt uses 20% of the sample for parameter tuning.

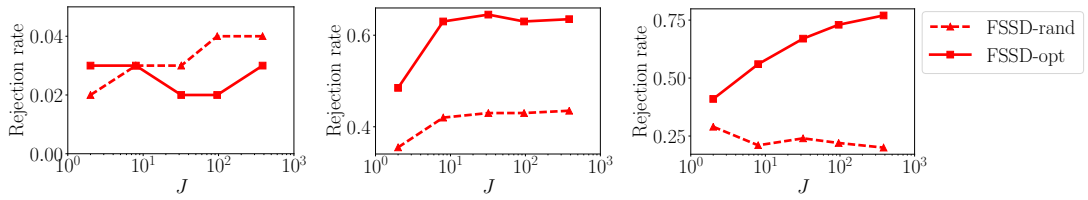
Figure 5.3d shows the rejection rates of all the tests as the sample size increases when p and q are the same RBM. All the tests have roughly the right false rejection rates at the set significance level $\alpha = 0.05$.

5.5.3 Informative Features

In the final simulation, we demonstrate that the learned test locations are informative in visualising where the model does not fit the data well. We consider crime data from the Chicago Police Department, recording $n = 11957$ locations (latitude-longitude coordinates) of robbery events in Chicago in 2016.[3] We address the situation in which a model p for the robbery location density is given, and we wish to visualise where it fails to match the data. We fit a Gaussian mixture model (GMM) with the expectation-maximization algorithm to a subsample of 5500 points. We then test the model on a held-out test set of the same size to obtain proposed locations of relevant features $\mathbf{v}$. Figure 5.5a shows the test robbery locations in purple, the model with two Gaussian components in wireframe, and the optimization objective for $\mathbf{v}$ as a grayscale contour plot (a red star indicates the maximum). We observe that the 2-component model is a poor fit to the data, particularly in the right tail areas of the data, as indicated in dark gray (i.e., the objective is high). Figure 5.5b shows a similar plot with a 10-component GMM. The additional components appear to have eliminated some mismatch in the right tail, however a discrepancy still exists in the left region. Here, the data have a sharp boundary on the right side following the geography of Chicago, and do not exhibit exponentially decaying Gaussian-like tails. We note that tests based on a learned feature located at the maximum both correctly reject H_0.

[3]Data can be found at `https://data.cityofchicago.org`.

5.5.4 Rejection Rate Vs. Number J of Test Locations



(a) SG. $d = 5$. $\alpha = 0.05$ (b) Gaussian vs. GMM. $d = 1$. (c) GVD. $d = 5$.

Figure 5.6: Plots of rejection rate against the number of test locations J in the three toy problems in Section 5.5.4.

The aim of this section is to explore the test power of the proposed FSSD test as a function of the number of test locations J. We consider three synthetic problems to illustrate three phenomena depending on the characteristic of the problem. We note that the test power may not necessarily increase with J. Figure 5.6 shows the rejection rate as a function of the test locations J in the three problems described below. In all cases, the sample size is set to $n = 500$, the train/test ratio is 50%, and the significance level is $\alpha = 0.05$. All rejection rates are computed with 200 trials with data sampled from the specified q in every trial.

We emphasize that the FSSD test is not designed to be used with large J, since doing so defeats the purpose of a linear-time test.

Same Gaussian (SG): In this problem, $p = q = \mathcal{N}(\mathbf{0}, \mathbf{I})$ in $\mathbb{R}^5$ i.e., H_0 is true. It can be seen in Figure 5.6a that both the FSSD tests with and without optimization achieve correct false positive rate at roughly α for all J considered. That is, under H_0, the false rejection rate stays at the right level for all J.

Gaussian vs. Gaussian Mixture Model (GMM): This is a one-dimensional problem where $p = \mathcal{N}(0, 1)$ and $q = 0.9\mathcal{N}(0, 1) + 0.1\mathcal{N}(0, 0.1^2)$ i.e., a mixture of two normal distributions. In this problem, p significantly differs from q in a small region around 0. This difference is created by the second mixture component. The characteristic of this problem is the local difference of p and q.

Figure 5.6b indicates that using random test locations (FSSD-rand) does not give high test power. With optimization (FSSD-opt), the power increases as J increases up to a point, after which it slightly drops down and reaches a plateau. This behavior can be explained by noting that there is only a very small region around 0 to detect the difference. More signal can be gained with diminishing return by increasing the number of test locations around 0. When J is sufficiently high, the increase in the variance of the statistic outweighs the gain of the signal (recall that the variance of the null distribution increases with J). This increase in the variance reduces the test power.

Gaussian Variance Difference (GVD): This is a synthetic problem studied in Jitkrittum et al. [2016] where $p = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $q = \mathcal{N}(\mathbf{0}, \mathrm{diag}(2, 1 \ldots, 1))$ in $\mathbb{R}^5$. In this case,

the region of difference between q and p exists only along the first dimension, and is broad.

In this case, Figure 5.6c shows that, with optimization, the power increases as the number of test locations increases. Unlike the case of Gaussian vs. GMM, the region of difference in this case is broad, and can accommodate more test locations to increase the signal. Despite this, we expect the test power to reach a plateau when J is sufficiently large for the same reason as described previously. In FSSD-rand, random test locations decrease the power due to the increase in the variance. Since only one dimension is relevant in determining the difference of p and q, it is unlikely that random locations are in the right region.

5.6 Known Results

This section presents known results that we use in our proofs.

Lemma 5.15 (Weyl's Perturbation Theorem [Bhatia, 2013, p. 152])**.** *Let $\lambda_j(A)$ denote the j^{th} eigenvalue of a square matrix A. If A, B are two Hermitian matrices, then*

$$\max_j |\lambda_j(A) - \lambda_j(B)| \leq \|A - B\|,$$

where $\|\cdot\|$ denotes the operator norm.

Lemma 5.16 ([van der Vaart, 2000, Lemma 21.2])**.** *For any sequence of cumulative distribution functions, $F_n^{-1} \xrightarrow{d} F^{-1}$ if and only if $F_n \xrightarrow{d} F$.*

Chapter 6

Informative Features for Automated Expectation Propagation

Summary We propose an efficient nonparametric strategy for learning a message operator in expectation propagation (EP), which takes as input the set of incoming messages to a factor node, and produces an outgoing message as output. This learned operator replaces the multivariate integral required in classical EP, which may not have an analytic expression. We use kernel-based regression, which is trained on a set of probability distributions representing the incoming messages, and the associated outgoing messages. The kernel approach has two main advantages: first, it is fast, as it is implemented using a novel two-layer random feature representation of the input message distributions; second, it has principled uncertainty estimates, and can be cheaply updated online, meaning it can request and incorporate new training data when it encounters inputs on which it is uncertain. In experiments, our approach is able to solve learning problems where a single message operator is required for multiple, substantially different data sets (logistic regression for a variety of classification problems), where it is essential to accurately assess uncertainty and to efficiently and robustly update the message operator.

6.1 Introduction

An increasing priority in Bayesian modelling is to make inference accessible and implementable for practitioners, without requiring specialist knowledge. This is a goal sought, for instance, in probabilistic programming languages [Wingate et al., 2011, Goodman et al., 2008], as well as in more granular, component-based systems [Stan Development Team, 2014, Minka et al., 2014]. In all cases, the user should be able to freely specify what they wish their model to express, without having to deal with the complexities of sampling, variational approximation, or distribution conjugacy. In reality, however, model convenience and simplicity can limit or undermine intended models, sometimes in ways the users might not expect. To take one example, the inverse gamma prior, which is widely used as a convenient conjugate prior for the variance, has quite pathological behaviour [Gelman, 2006]. In general, more expressive,

freely chosen models are more likely to require expensive sampling or quadrature approaches, which can make them challenging to implement or impractical to run.

We address the particular setting of expectation propagation [Minka, 2001], a message passing algorithm wherein messages are confined to being members of a particular parametric family. The process of integrating incoming messages over a factor potential, and projecting the result onto the required output family, can be difficult, and in some cases not achievable in closed form. Thus, a number of approaches have been proposed to implement EP updates numerically, independent of the details of the factor potential being used. One approach, due to Barthelmé and Chopin [2011], is to compute the message update via importance sampling. While these estimates converge to the desired integrals for a sufficient number of importance samples, the sampling procedure must be run at every iteration during inference, hence it is not viable for large-scale problems.

An improvement on this approach is to use importance sampled instances of input/output message pairs to train a regression algorithm, which can then be used in place of the sampler. Heess et al. [2013] use neural networks to learn the mapping from incoming to outgoing messages, and the learned mappings perform well on a variety of practical problems. This approach comes with a disadvantage: it requires training data that cover the entire set of possible input messages for a given type of problem (e.g., datasets representative of all classification problems the user proposes to solve), and it has no way of assessing the uncertainty of its prediction, or of updating the model online in the event that a prediction is uncertain.

The disadvantages of the neural network approach were the basis for work by Eslami et al. [2014], who replaced the neural networks with random forests. The random forests provide uncertainty estimates for each prediction. This allows them to be trained "just-in-time," during EP inference, whenever the predictor decides it is uncertain. Uncertainty estimation for random forests relies on unproven heuristics, however: we demonstrate empirically that such heuristics can become highly misleading as we move away from the initial training data. Moreover, online updating can result in unbalanced trees, resulting in a cost of prediction of $\mathcal{O}(N)$ for training data of size N, rather than the ideal $\mathcal{O}(\log(N))$.

Proposal We propose a novel, kernel-based approach to learning a message operator nonparametrically for expectation propagation. The learning algorithm takes the form of a distribution regression problem [Szabó et al., 2016, Oliva et al., 2013, Poczos et al., 2013], where the inputs are probability measures represented as embeddings of the distributions to a reproducing kernel Hilbert space (RKHS), and the outputs are vectors of message parameters. A first advantage of this approach is that one does not need to pre-specify customized features of the distributions, as in Eslami et al. [2014], Heess et al. [2013]. Rather, we use a general characteristic kernel on input distributions [Christmann and Steinwart, 2010, Eq. 9], which in our experiments gives better performance than customized features. A potential downside of the

kernel approach is that it can be computationally costly, with training time of $\mathcal{O}(N^3)$ and a cost of $\mathcal{O}(N)$ to make a prediction. To make the algorithm computationally tractable, we regress directly in the primal from random Fourier features of the data [Rahimi and Recht, 2007, Le et al., 2013, Yang et al., 2015]. In particular, we establish a novel random feature representation for when inputs are distributions, via a two-level random feature approach. This gives us both fast prediction (linear in the number of random features), and fast online updates (quadratic in the number of random features).

A second advantage of our approach is that, being an instance of Gaussian process regression, there are well established estimates of predictive uncertainty [Rasmussen and Williams, 2006, Ch. 2]. We use these uncertainty estimates so as to determine when to query the importance sampler for additional input/output pairs, i.e., the predictive uncertainty triggers just-in-time updates of the regressor. We demonstrate empirically that our uncertainty estimates are more robust and informative than those for random forests, especially as we move away from the training data.

In Section 6.2, we introduce the notation for expectation propagation, and indicate how an importance sampling procedure can be used as an oracle to provide training data for the message operator. Next, in Section 6.3, we describe our kernel regression approach, and the form of an efficient kernel message operator mapping the input messages (distributions embedded in an RKHS) to outgoing messages (sets of parameters of the outgoing messages). Finally, in Section 6.4, we describe our experiments, which cover three topics: a benchmark of our uncertainty estimates, a demonstration of factor learning on artificial data with well-controlled statistical properties, and a logistic regression experiment on four different real-world datasets, demonstrating that our just-in-time learner can correctly evaluate its uncertainty and update the regression function as the incoming messages change.

6.2 Message Passing

Message passing is a family of algorithms for computing marginal distributions over a subset of variables, given a joint distribution. Assume that the joint probability density p over a set of variables $\mathbf{x} = (x_1, \dots x_d)$ of interest can be represented as a product of J factors i.e., $p(\mathbf{x}) = \frac{1}{Z}\prod_{j=1}^{J} f_j(\mathbf{x}_{\mathrm{ne}(f_j)})$, and Z is the normalization constant [Bishop, 2006, Section 8.4.3: Factor Graphs]. For each $j \in \{1, \dots, J\}$, the factors f_j is a non-negative function (not necessarily a probability density) defined over a subset $\mathbf{x}_{\mathrm{ne}(f_j)}$ of the full set of variables $\mathbf{x}$, where $\mathrm{ne}(f_j) \subset \{1, \dots, d\}$ denotes the set of indices of variables connected to f_j. These variables form the neighbors of the factor node f_j when $p(\mathbf{x})$ is represented as a factor graph. An example of a simple probability density $p(\mathbf{x})$ and its factor graph representation is shown in Figure 6.1.

We deal with models in which some of the factors have a non-standard form, or may not have a known closed-form expression (i.e., black-box factors). Although our approach applies to any such factor in principle, in this paper we focus on *directed*

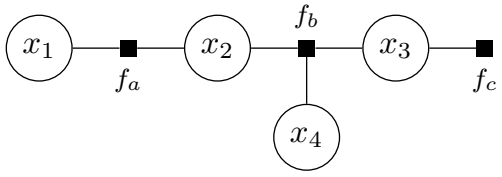


Figure 6.1: An example of a simple factor graph corresponding to $p(\mathbf{x}) = \frac{1}{Z} f_a(x_1, x_2) f_b(x_2, x_3, x_4) f_c(x_3)$, containing three factors f_a, f_b, f_c defined over four variables $x_1, \ldots, x_4$.

factors $f(\mathbf{x}_{\text{out}}|\mathbf{x}_{\text{in}})$ which specify a conditional distribution over variables $\mathbf{x}_{\text{out}}$ given $\mathbf{x}_{\text{in}}$ (and thus $\mathbf{x}_{\text{ne}(f)} = (\mathbf{x}_{\text{out}}, \mathbf{x}_{\text{in}})$). The only assumption we make is that we are provided with a forward sampling function $f : \mathbf{x}_{\text{in}} \mapsto \mathbf{x}_{\text{out}}$, i.e., a function that maps (stochastically or deterministically) a setting of the input variables $\mathbf{x}_{\text{in}}$ to a sample from the conditional distribution over $\mathbf{x}_{\text{out}} \sim f(\cdot|\mathbf{x}_{\text{in}})$. In particular, the ability to evaluate the potential value of $f(\mathbf{x}_{\text{out}}|\mathbf{x}_{\text{in}})$ is not assumed. A natural way to specify f is as code in a probabilistic program.

6.2.1 Belief Propagation and Expectation Propagation

Belief propagation (BP), or the sum-product message passing [Pearl, 1982], computes *exact* marginal distributions over subsets of variables (possibly conditioned on another subset) by iteratively passing messages between variables and factors. A BP message sent from a factor f to variable x_i (where $i \in \text{ne}(f)$) is given by

$$m_{f \to i}(x_i) = \int f(\mathbf{x}_{\text{ne}(f)}) \prod_{i' \in \text{ne}(f) \setminus \{i\}} m_{i' \to f}(x_{i'}) \, \mathrm{d}\mathbf{x}_{\text{ne}(f) \setminus \{i\}} \tag{6.1}$$

where $m_{i' \to f}$ is the message sent to factor f from a neighboring variable $x_{i'}$.[1] A factor-to-variable message in general can be any non-negative function (i.e., not necessarily a distribution) whose complexity depends on the factor f. A variable-to-factor message $m_{i' \to f}$ is simply the product of all messages from all other neighboring factors (except the recipient factor):

$$m_{i' \to f}(x_{i'}) = \prod_{f' \in \text{ne}(i') \setminus \{f\}} m_{f' \to x_{i'}}(x_{i'}), \tag{6.2}$$

where $\text{ne}(i')$ is the set of neighboring factor nodes to the variable i'. The marginal distribution $p(x_k)$ of x_k is proportional to the product of all incoming messages to x_k:

$$p(x_k) \propto \prod_{f' \in \text{ne}(k)} m_{f' \to x_k}(x_k). \tag{6.3}$$

If a variable is observed, conditioning on the observed value amounts to replacing the incoming message from the variable with the Dirac delta function centered at the observed value. It is well known that if the factor graph is a tree, BP ensures consistency

[1] We will interchangeably write $m_{i \to f}$ and $m_{x_i \to f}$ for a message from the variable x_i to f. Similarly, we use both $m_{f \to i}$ and $f_{f \to x_i}$ to denote a message from the factor f to x_i.

of the obtained marginals at convergence [Wainwright and Jordan, 2008, p. 105]. That is, the marginal computed in (6.3) is equal to $p(x_k) = \int \frac{1}{Z} \prod_{j=1}^{J} f_j(\mathbf{x}_{\text{ne}(f_j)}) \, \mathrm{d}\mathbf{x}_{\setminus k}$ where $\mathbf{x}_{\setminus k} = (x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_d)$.

6.2.2 Expectation Propagation

Expectation Propagation (EP) is an *approximate* procedure for computing marginal beliefs of variables by iteratively passing messages between variables and factors until convergence [Minka, 2001]. It can be seen as an alternative to belief propagation, where a part of factor outgoing messages is projected onto a prespecified class $\mathcal{Q}$ of known parametric distributions. In contrast to BP, the projection in EP allows messages to be represented parametrically regardless of the complexity of the factors in the graph. The EP message from factor f to variable $V \in \text{ne}(f)$ is

$$m_{f\to V}(x_V) = \frac{\text{proj}[\frac{1}{Z} \int f(\mathbf{x}_{\text{ne}(f)}) \prod_{V' \in \text{ne}(f)} m_{V'\to f}(x_{V'}) \, \mathrm{d}\mathbf{x}_{\text{ne}(f)\setminus V}]}{m_{V\to f}(x_V)} := \frac{q_{f\to V}(x_V)}{m_{V\to f}(x_V)}, \quad (6.4)$$
$$\text{proj}[p] := \text{argmin}_{q\in\mathcal{Q}} \text{KL}[p||q],$$

where $\text{KL}[p||q]$ denotes the Kullback-Leibler (KL) divergence from p to q, Z is the normalizing constant, and $\mathcal{Q}$ is a chosen class of distributions, typically in the exponential family e.g., the set of normal distributions. EP is a general case of BP where factor-outgoing messages are defined such that the marginal of each variable x_V is constrained to be in $\mathcal{Q}$. To see this, consider the argument to the proj operator in (6.4):

$$\begin{aligned}
&\int f(\mathbf{x}_{\text{ne}(f)}) \prod_{V' \in \text{ne}(f)} m_{V'\to f}(x_{V'}) \, \mathrm{d}\mathbf{x}_{\text{ne}(f)\setminus V} \\
&= m_{V\to f}(x_V) \int f(\mathbf{x}_{\text{ne}(f)}) \prod_{V' \in \text{ne}(f)\setminus\{V\}} m_{V'\to f}(x_{V'}) \, \mathrm{d}\mathbf{x}_{\text{ne}(f)\setminus V} \qquad (6.5) \\
&\overset{(a)}{=} m_{V\to f}(x_V) m_{f\to V}(x_V) \\
&\overset{(b)}{=} \left(\prod_{f' \in \text{ne}(V)\setminus\{f\}} m_{f'\to V}(x_V) \right) m_{f\to V}(x_V) \\
&\overset{(c)}{\propto} p(x_V),
\end{aligned}$$

where (a) follows from (6.1), at (b) we use (6.2), and at (c) follows from (6.3). If $p(x_V)$ is already in the chosen family $\mathcal{Q}$, then the projection in (6.4) has no effect, and EP factor-outgoing message $m_{f\to V}(x_V)$ reduces to the BP outgoing message in (6.1). An EP message from a variable is the same as the BP message in (6.2). The marginal $p(x_k)$ is proportional to the product of all EP incoming messages to x_k as in (6.3).

There are two key differences to BP. Firstly, the marginal estimated by EP messages is generally not exact because of the projection. Secondly, EP is inherently loopy, and requires multiple passes over all the nodes in the graph, even when the graph is a tree. This characteristic is due to the cyclic dependency on the variable-to-factor

message $m_{V\to f}$, in the computation of the factor-to-variable message $m_{f\to V}$ (see (6.4)). In BP, if the graph is a tree, then there is no such cyclic dependency (see (6.1)), and messages can be sent from leaf nodes to the node whose marginal distribution is desired, passing over all the variable nodes only once. One way to implement EP is by initializing all variable-to-factor messages appropriately, and iteratively computing (6.4) until convergence i.e., until factor-to-variable messages no longer change.

6.2.3 Monte Carlo EP Message Approximation

A common choice for $\mathcal{Q}$ (for the projection in (6.4)) is to set to a subset of the exponential family. Consider a density $q(x|\eta) = h(x)\exp\left(\eta^\top u(x) - A(\eta)\right)$ in the exponential family parameterized by a natural parameter $\eta \in \mathbb{R}^T$, where $u(x)$ computes the sufficient statistics of x, $h(x)$ is the base measure, and $A(\eta) := \log \int h(x)\exp\left(\eta^\top u(x)\right)\,\mathrm{d}x$ is the log normalizer of q. Assume that $\mathcal{Q} := \left\{q(x|\eta) \mid \eta \in \mathbb{R}^T\right\}$ for some fixed u and h. With this choice for $\mathcal{Q}$, it is well known that projection amounts to finding a distribution in $\mathcal{Q}$ which satisfies a moment-matching constraint. Specifically, given an arbitrary distribution $r(x)$, its projection onto $\mathcal{Q}$ is given by

$$\begin{aligned}\mathrm{proj}[r] &= \arg\min_{q\in\mathcal{Q}} \mathrm{KL}[r||q] = q^*(x), \text{ such that}\\ \mathbb{E}_{x\sim r}[u(x)] &= \mathbb{E}_{x\sim q^*}[u(x)].\end{aligned} \tag{6.6}$$

We observe from (6.6) that finding the projection of $r(x)$ requires evaluating the expectation of $u(x)$ under r. Exactly computing the expectation can be challenging, as it requires evaluating a high-dimensional integral with respect to an arbitrary distribution r. Even closed-form factors often require hand-crafted approximations, or the use of expensive numerical integration techniques; for "black-box" factors implemented as forward sampling functions as in the problem we consider, more sophisticated techniques are needed. Barthelmé and Chopin [2011], Heess et al. [2013], Eslami et al. [2014] propose a stochastic approach to the integration and projection step. A sample based approximation of this expectation can be obtained via importance sampling described as follows.

Approximating Factor-Outgoing Messages with Importance Sampling Assume that $\mathbf{x}_{\mathrm{ne}(f)} = (\mathbf{x}_{\mathrm{out}}, \mathbf{x}_{\mathrm{in}})$ so that $f(\mathbf{x}_{\mathrm{ne}(f)}) = f(\mathbf{x}_{\mathrm{out}}|\mathbf{x}_{\mathrm{in}})$, and that $\mathbf{x}_{\mathrm{out}} \sim f(\cdot|\mathbf{x}_{\mathrm{in}})$ can be simulated given $\mathbf{x}_{\mathrm{in}}$. Let

$$r_{f\to V}(x_V) := \frac{1}{Z_r}\int f(\mathbf{x}_{\mathrm{out}}|\mathbf{x}_{\mathrm{in}}) \prod_{V'\in\mathrm{ne}(f)} m_{V'\to f}(x_{V'})\,\mathrm{d}\mathbf{x}_{\mathrm{ne}(f)\setminus V}$$

be the function that is the argument to the projection in (6.4), where Z_r is the normalizer of $r_{f\to V}$. Given a proposal distribution $s(\mathbf{x}_{\mathrm{in}})$ (i.e., a normalized probability density) with appropriate support, the expectation $\mathbb{E}_{x_V\sim r_{f\to V}}[u(x_V)]$ in (6.6) can be

approximated as

$$\begin{aligned}\mathbb{E}_{x_V \sim r_{f\to V}}[u(x_V)] &= \frac{1}{Z_r}\int\int u(x_V) f(\mathbf{x}_{\text{out}}|\mathbf{x}_{\text{in}}) \prod_{V'\in \text{ne}(f)} m_{V'\to f}(x_{V'})\, \mathrm{d}\mathbf{x}_{\text{ne}(f)\setminus V}\, \mathrm{d}x_V \\ &= \frac{1}{Z_r}\int\int u(x_V) \frac{\prod_{V'\in \text{ne}(f)} m_{V'\to f}(x_{V'})}{s(\mathbf{x}_{\text{in}})} f(\mathbf{x}_{\text{out}}|\mathbf{x}_{\text{in}}) s(\mathbf{x}_{\text{in}})\, \mathrm{d}\mathbf{x}_{\text{out}}\mathrm{d}\mathbf{x}_{\text{in}} \\ &\approx \frac{\frac{1}{M}\sum_{l=1}^{M} \tilde{w}(\mathbf{x}^{(l)}_{\text{ne}(f)}) u(x^{(l)}_V)}{\frac{1}{M}\sum_{j=1}^{M} \tilde{w}(\mathbf{x}^{(j)}_{\text{ne}(f)})}, \end{aligned} \tag{6.7}$$

where $\tilde{w}(\mathbf{x}_{\text{ne}(f)}) = \frac{\prod_{V'\in \text{ne}(f)} m_{V'\to f}(x_{V'})}{s(\mathbf{x}_{\text{in}})}$ is the importance weight, and $\{\mathbf{x}^{(l)}_{\text{ne}(f)}\}_{l=1}^{M} \sim f(\mathbf{x}_{\text{out}}|\mathbf{x}_{\text{in}}) s(\mathbf{x}_{in})$ are M Monte Carlo particles used to approximate the expectation. We note that

$$\begin{aligned} Z_r &= \int f(\mathbf{x}_{\text{out}}|\mathbf{x}_{\text{in}}) \prod_{V'\in \text{ne}(f)} m_{V'\to f}(x_{V'})\, \mathrm{d}\mathbf{x}_{\text{ne}(f)} \\ &= \int \frac{\prod_{V'\in \text{ne}(f)} m_{V'\to f}(x_{V'})}{s(\mathbf{x}_{\text{in}})} f(\mathbf{x}_{\text{out}}|\mathbf{x}_{\text{in}}) s(\mathbf{x}_{\text{in}})\, \mathrm{d}\mathbf{x}_{\text{ne}(f)} \\ &\approx \frac{1}{M}\sum_{j=1}^{M} \tilde{w}(\mathbf{x}^{(j)}_{\text{ne}(f)}). \end{aligned}$$

The estimate for the expected sufficient statistics in (6.7) only assumes that the proposal distribution $s(\mathbf{x}_{\text{in}})$ is normalized. In particular, we assume neither the ability to evaluate the density of $f(\mathbf{x}_{\text{out}}|\mathbf{x}_{\text{in}})$, nor the fact that $\prod_{V'\in \text{ne}(f)} m_{V'\to f}(x_{V'})$ is a normalized density. The moment estimate in (6.7) provides us with an estimate of the the natural parameter η of $q_{f\to V}$ in (6.4), from which the factor-outgoing message $m_{f\to V}$ is readily computed.

6.2.4 Learning to Pass EP Messages

Message approximation as in the previous section could be used directly when running the EP algorithm, as in Barthelmé and Chopin [2011]; but this approach can suffer when the number of particles M is small, and the importance sampling estimate is not reliable. On the other hand, for large M, the computational cost of running EP with approximate messages can be very high, as importance sampling must be performed for sending each outgoing message. To obtain a low-variance message approximation at lower computational cost, Heess et al. [2013] and Eslami et al. [2014] both amortize previously computed approximate messages by training a function approximator to directly map a tuple of incoming variable-to-factor messages $(m_{V'\to f})_{V'\in \text{ne}(f)}$ to an approximate factor-to-variable message $m_{f\to V}$, i.e., they learn a mapping

$$M^{\theta}_{f\to V} : (m_{V'\to f})_{V'\in \text{ne}(f)} \mapsto m_{f\to V},$$

where θ is a tunable parameter vector. We will refer to $M^{\theta}_{f\to V}$ as a *message operator*, $(m_{V'\to f})_{V'\in \text{ne}(f)}$ as incoming messages, and $m_{f\to V}$ as outgoing message.

Heess et al. [2013] use neural networks and a large, fixed training set to learn their approximate message operator prior to running EP. By contrast, Eslami et al. [2014] employ random forests as their class of learning functions, and update their approximate message operator on the fly during inference (known as just-in-time learning), depending on the predictive uncertainty of the current message operator. Specifically, they endow their function approximator with an uncertainty estimate

$$\mathfrak{v}^{\theta}_{f\to V} : (m_{V'\to f})_{V'\in \mathrm{ne}(f)} \mapsto \mathfrak{v},$$

where $\mathfrak{v}$ indicates the expected unreliability of the predicted, approximate message $m_{f\to V}$ returned by $M^{\theta}_{f\to V}$. If $\mathfrak{v} = \mathfrak{v}^{\theta}_{f\to V}\left((m_{V'\to f})_{V'\in \mathrm{ne}(f)}\right)$ exceeds a pre-defined threshold, the required message is approximated via importance sampling (as in (6.7)), and the message operator $M^{\theta}_{f\to V}$ is updated online with this new labeled pair $\left((m_{V'\to f})_{V'\in \mathrm{ne}(f)}, m_{f\to V}\right)$ leading to a new set of parameters θ' with

$$\mathfrak{v}^{\theta'}_{f\to V}\left((m_{V'\to f})_{V'\in \mathrm{ne}(f)}\right) < \mathfrak{v}^{\theta}_{f\to V}\left((m_{V'\to f})_{V'\in \mathrm{ne}(f)}\right).$$

Eslami et al. [2014] estimate the predictive uncertainty $\mathfrak{v}^{\theta}_{f\to V}$ via the heuristic of looking at the variability of the predictions from all the trees in the forest [Criminisi and Shotton, 2013]. They implement their online updates by splitting the trees at their leaves. Both these mechanisms can be problematic, however. First, the heuristic used in computing uncertainty has no guarantees: indeed, uncertainty estimation for random forests remains a challenging topic of current research [Hutter, 2009]. This is not merely a theoretical consideration: in our experiments in Section 6.4, we demonstrate that uncertainty heuristics for random forests become unstable and inaccurate as we move away from the initial training data. Second, online updates of random forests may not work well when the newly observed data are from a very different distribution to the initial training sample [e.g. Lakshminarayanan et al., 2014, Fig. 3]. For large amounts of training set drift, the leaf-splitting approach of Eslami et al. can result in a decision tree in the form of a long chain, giving a worst case cost of prediction (computational and storage) of $O(N)$ for training data of size N, vs the ideal of $O(\log(N))$ for balanced trees. Finally, note that the approach of Eslami et al. uses certain bespoke features of the factors when specifying tree traversal in the random forests, notably the value of the factor potentials at the mean and mode of the incoming messages. These features require expert knowledge of the model on the part of the practitioner, and are not available in the "forward sampling" setting. The present work does not employ such features.

Computational Complexity of the Random Forest Approach We now provide a cost breakdown of the random forest approach. Let K be the number of trees in the random forest, D_t be the number of features used in tree traversal, D_r be the number of features used in making predictions at the leaves, N be the number of training points provided by the importance sampling oracle, and L be the number of training

points per leaf. Assuming that the depth of trees is $\mathcal{O}(\log(N))$, one prediction from a random forest costs $\mathcal{O}(KD_rD_t\log(N))$, and one update costs $\mathcal{O}(KD_r^3D_t\log(N))$. This is because for each of K trees in the forest, tree traversal involves $\mathcal{O}(\log(N))$ steps which each costs $\mathcal{O}(D_t)$ (splitting on an internal node involves a linear regression using the tree traversal features), and one prediction costs $\mathcal{O}(D_r)$. The leaf predictions are made using polynomial regression of degree two, which must be re-trained from scratch for each new point. Retraining at a leaf costs $\mathcal{O}(D_r^2(D_r+L))$, however L is typically negligible at a lower depth, effectively costing $\mathcal{O}(D_r^3)$. Therefore, the total updating cost across all trees is $\mathcal{O}(KD_r^3D_t\log(N))$.

It is instructive to consider some representative numbers used by Eslami et al. The number of trees in the forest was $K = 64$, the number of features at a leaf was in general $D_r = 14$ (since Gaussian, Beta and Gamma distributions are all parameterised by two numbers, most factors had two incoming messages, and quadratic regressors were used). D_t was typically of the order of 10 to 20, depending on the number of incoming factor messages: these include parameters of the messages, the values of the factor at the mean and mode of the incoming messages, and the binary features characterizing the message. Training set size N was in the order of thousands (1,000 to 5,000), and the number of samples L per leaf was between 10 and 50.

6.3 Proposal: Kernel-Based Message Operators

We now propose a kernel regression method for jointly learning the message operator $M^{\theta}_{f\to V}$ and uncertainty estimate $\mathfrak{v}^{\theta}_{f\to V}$. We regress from the tuple of incoming messages, which are probability distributions, to the parameters of the outgoing message. To this end we apply a kernel over distributions from [Christmann and Steinwart, 2010] to the case where the input consists of more than one distribution. We note that Song et al. [2010, 2011] propose a related regression approach for predicting outgoing messages from incoming messages, for the purpose of belief propagation. Their setting is different from ours, however, as their messages are smoothed conditional density functions rather than parametric distributions of known form. An important issue is computational cost, as EP is an iterative algorithm and many predictions must be made. To achieve fast predictions and message operator updates, we follow Rahimi and Recht [2007], Le et al. [2013], Yang et al. [2015], and express the kernel regression in terms of random features whose expected inner product is equal to the kernel function; i.e., we perform regression directly in the primal on these random features. In Section 6.3.1, we define our kernel on tuples of distributions, and then derive the corresponding random feature representation in Section 6.3.2. Section 6.3.3 describes the regression algorithm, as well as our strategy for uncertainty evaluation and online updates.

6.3.1 Kernels on Tuples of Distributions

In the followings, we consider only a single factor, and therefore drop the factor identity from our notation. We write the set of c incoming messages to a factor node

as a tuple of probability densities $R := (r^{(l)})_{l=1}^{c}$ of random variables $X^{(l)}$ on respective domains $\mathcal{X}^{(l)}$. Our goal is to define a kernel between one such tuple, and a second one, which we will write $S := (s^{(l)})_{l=1}^{c}$.

We define our kernel in terms of embeddings of the tuples R, S into a reproducing kernel Hilbert space (RKHS). We first consider the embedding of a single distribution in the tuple. Let us define an RKHS $\mathcal{H}^{(l)}$ on each domain, with respective kernel $k^{(l)}(x_1^{(l)}, x_2^{(l)})$ on $\mathcal{X}^{(l)} \times \mathcal{X}^{(l)}$. We may embed individual probability distributions to these RKHSs (see Section 2.2: Kernel Mean Embedding). The mean embedding of $r^{(l)}$ is written

$$\mu_{r^{(l)}}(\cdot) := \int k^{(l)}(x^{(l)}, \cdot) r^{(l)}(x^{(l)}) \, \mathrm{d}x^{(l)}.$$

Similarly, a mean embedding may be defined on the product of messages in a tuple $\mathsf{r} = \times_{l=1}^{c} r^{(l)}$ as

$$\mu_{\mathsf{r}} := \int k([x^{(1)}, \ldots, x^{(c)}], \cdot) \mathsf{r}(x^{(1)}, \ldots, x^{(c)}) \, \mathrm{d}x^{(1)} \cdots \mathrm{d}x^{(c)}, \tag{6.8}$$

where we have defined the joint kernel k on the product space $\mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(c)}$, associated with the RKHS $\mathcal{H} = \mathcal{H}^{(1)} \times \cdots \times \mathcal{H}^{(c)}$. Finally, a kernel on two such embeddings $\mu_{\mathsf{r}}, \mu_{\mathsf{s}}$ of tuples R, S can be obtained as in Christmann and Steinwart [2010, eq. 9],

$$\kappa(\mathsf{r}, \mathsf{s}) = \exp\left(-\frac{\|\mu_{\mathsf{r}} - \mu_{\mathsf{s}}\|_{\mathcal{H}}^2}{2\gamma^2}\right). \tag{6.9}$$

This kernel has two parameters: γ^2, and the width parameter of the kernel k defining $\mu_{\mathsf{r}} = \mathbb{E}_{\mathbf{x} \sim \mathsf{r}} k(\mathbf{x}, \cdot)$.

We have considered several alternative kernels on tuples of messages, including kernels on the message parameters, kernels on a tensor feature space of the distribution embeddings in the tuple, and inner products of the mean embeddings (6.8). We have found these alternatives to have worse empirical performance than the approach described above. We give details of these experiments in Section 6.7.

6.3.2 Random Feature Approximations

One approach to learning the mapping $M_{f \to V}^{\theta}$ from incoming to outgoing messages would be to employ Gaussian process regression [Rasmussen and Williams, 2006], using the kernel in (6.9). This approach is not suited to just-in-time (JIT) learning, however, as both prediction and storage costs grow with the size of the training set; thus, inference on even moderately sized datasets rapidly becomes computationally prohibitive. Instead, we define a finite-dimensional random feature map $\hat{\psi} \in \mathbb{R}^{D_{\text{out}}}$ such that $\kappa(\mathsf{r}, \mathsf{s}) \approx \hat{\psi}(\mathsf{r})^\top \hat{\psi}(\mathsf{s})$, and regress directly on these feature maps in the primal (see next section). Storage and computation are then a function of the dimension of the feature map D_{out}, yet performance is close to that obtained using the exact kernel κ.

In Rahimi and Recht [2007], a method based on Fourier transforms was proposed for computing a vector of random features $\hat{\varphi}$ for a translation invariant kernel $k(\mathbf{x}, \mathbf{y}) =$

Algorithm 6.1 Construction of two-stage random features for κ

Require: Input distribution r, Fourier transform $\hat{k}$ of the embedding translation-invariant kernel k, number of inner features D_{in}, number of outer features D_{out}, outer Gaussian width γ^2.

Ensure: Random features $\hat{\psi}(\mathsf{r}) \in \mathbb{R}^{D_{\text{out}}}$.

1: Sample $\{\omega_i\}_{i=1}^{D_{\text{in}}} \overset{i.i.d}{\sim} \hat{k}$.

2: Sample $\{b_i\}_{i=1}^{D_{\text{in}}} \overset{i.i.d}{\sim} \text{Uniform}[0, 2\pi]$.

3: $\hat{\phi}(\mathsf{r}) = \sqrt{\frac{2}{D_{\text{in}}}} \left(\mathbb{E}_{\mathbf{x}\sim\mathsf{r}} \cos(\omega_i^\top \mathbf{x} + b_i)\right)_{i=1}^{D_{\text{in}}} \in \mathbb{R}^{D_{\text{in}}}$
If $\mathsf{r}(\mathbf{x}) = \mathcal{N}(\mathbf{x}; m, \Sigma)$,

$$\hat{\phi}(\mathsf{r}) = \sqrt{\frac{2}{D_{\text{in}}}} \left(\cos(\omega_i^\top m + b_i) \exp\left(-\frac{1}{2}\omega_i^\top \Sigma \omega_i\right)\right)_{i=1}^{D_{\text{in}}}.$$

4: Sample $\{\nu_i\}_{i=1}^{D_{\text{out}}} \overset{i.i.d}{\sim} \hat{k}_{\text{gauss}}(\gamma^2)$ i.e., Fourier transform of a Gaussian kernel with width γ^2.

5: Sample $\{c_i\}_{i=1}^{D_{\text{out}}} \overset{i.i.d}{\sim} \text{Uniform}[0, 2\pi]$.

6: $\hat{\psi}(\mathsf{r}) = \sqrt{\frac{2}{D_{\text{out}}}} \left(\cos(\nu_i^\top \hat{\phi}(\mathsf{r}) + c_i)\right)_{i=1}^{D_{\text{out}}} \in \mathbb{R}^{D_{\text{out}}}$

$k(\mathbf{x} - \mathbf{y})$ such that $k(\mathbf{x}, \mathbf{y}) \approx \hat{\varphi}(\mathbf{x})^\top \hat{\varphi}(\mathbf{y})$ where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\hat{\varphi}(\mathbf{x}), \hat{\varphi}(\mathbf{y}) \in \mathbb{R}^{D_{\text{in}}}$. This is possible because of Bochner's theorem [Rudin, 2013], which states that a continuous, translation-invariant kernel k can be written in the form of an inverse Fourier transform:

$$k(\mathbf{x} - \mathbf{y}) = \int \hat{k}(\omega) e^{i\omega^\top (\mathbf{x}-\mathbf{y})} \, \mathrm{d}\omega,$$

where $i = \sqrt{-1}$ and the Fourier transform $\hat{k}$ of the kernel can be treated as a distribution. The inverse Fourier transform can thus be seen as an expectation of the complex exponential, which can be approximated with a Monte Carlo average by drawing random frequencies from $\hat{k}$. We will follow a similar approach, and derive a two-stage set of random Fourier features for (6.9). Details of the random features of Rahimi and Recht [2007] are given in Section 6.6.

We start by expanding the exponent of (6.9) as

$$\kappa(\mathsf{r}, \mathsf{s}) = \exp\left(-\frac{1}{2\gamma^2}\langle \mu_\mathsf{r}, \mu_\mathsf{r} \rangle + \frac{1}{\gamma^2}\langle \mu_\mathsf{r}, \mu_\mathsf{s} \rangle - \frac{1}{2\gamma^2}\langle \mu_\mathsf{s}, \mu_\mathsf{s} \rangle\right). \tag{6.10}$$

Assume that the embedding kernel k used to define the embeddings μ_r and μ_s is translation invariant. Since $\langle \mu_\mathsf{r}, \mu_\mathsf{s} \rangle = \mathbb{E}_{x\sim\mathsf{r}}\mathbb{E}_{y\sim\mathsf{s}} k(\mathbf{x} - \mathbf{y})$, one can use the result of Rahimi and Recht [2007] to write

$$\begin{aligned}\langle \mu_\mathsf{r}, \mu_\mathsf{s} \rangle &\approx \mathbb{E}_{\mathbf{x}\sim\mathsf{r}}\mathbb{E}_{\mathbf{y}\sim\mathsf{s}} \hat{\varphi}(\mathbf{x})^\top \hat{\varphi}(\mathbf{y}) \\ &= \mathbb{E}_{\mathbf{x}\sim\mathsf{r}} \hat{\varphi}(\mathbf{x})^\top \mathbb{E}_{\mathbf{y}\sim\mathsf{s}} \hat{\varphi}(\mathbf{y}) := \hat{\phi}(\mathsf{r})^\top \hat{\phi}(\mathsf{s}),\end{aligned} \tag{6.11}$$

where the output of the map $\hat{\phi}(\mathsf{r}) \colon \mathsf{r} \mapsto \mathbb{E}_{\mathbf{x}\sim\mathsf{r}} \hat{\varphi}(\mathbf{x})$ contains D_{in} standard Rahimi-Recht random features, shown in Steps 1-3 of Algorithm 6.1. By combining (6.10) and (6.11),

we have

$$\kappa(\mathsf{r},\mathsf{s}) \approx \exp\left(-\frac{\|\hat{\phi}(\mathsf{r})-\hat{\phi}(\mathsf{s})\|_{D_{\text{in}}}^2}{2\gamma^2}\right),$$

which is a standard Gaussian kernel on $\mathbb{R}^{D_{\text{in}}}$, and $\|\cdot\|_{D_{\text{in}}}$ denotes the standard Euclidean norm in $\mathbb{R}^{D_{\text{in}}}$. We can thus further approximate this Gaussian kernel by the random Fourier features of Rahimi and Recht, to obtain a vector of random features $\hat{\psi}$ such that $\kappa(\mathsf{r},\mathsf{s}) \approx \hat{\psi}(\mathsf{r})^\top \hat{\psi}(\mathsf{s})$ where $\hat{\psi}(\mathsf{r}), \hat{\psi}(\mathsf{s}) \in \mathbb{R}^{D_{\text{out}}}$. Pseudocode for generating the random features $\hat{\psi}$ is given in Algorithm 6.1. Note that the sine component in the complex exponential vanishes due to the translation invariance property (analogous to an even function), i.e., only the cosine term remains. We refer to Section 6.6.3 for more details on the derivation.

For the implementation, we need to pre-compute $\{\omega_i\}_{i=1}^{D_{\text{in}}}, \{b_i\}_{i=1}^{D_{\text{in}}}, \{\nu_i\}_{i=1}^{D_{\text{out}}}$ and $\{c_i\}_{i=1}^{D_{\text{out}}}$, where D_{in} and D_{out} are the number of random features used. A more storage-efficient way to support a large number of random features is to store only the random seed used to generate the features, and to generate the coefficients on-the-fly when needed [Dai et al., 2014]. In our implementation, we use a Gaussian kernel for k.

6.3.3 Regression for Message Prediction

Let $\mathsf{X} = (\mathsf{x}_1|\cdots|\mathsf{x}_N)$ be the N training tuples of incoming messages to a factor node, represented by the two-stage random features described in Section 6.3.2 i.e., $\mathsf{x}_i \in \mathbb{R}^{D_{\text{out}}}$ for $i = 1,\ldots,N$. Let $\mathsf{Y} = \left(\mathbb{E}_{x_V\sim q^1_{f\to V}} u(x_V)|\cdots|\mathbb{E}_{x_V\sim q^N_{f\to V}} u(x_V)\right) \in \mathbb{R}^{D_y\times N}$ be the expected sufficient statistics of the corresponding outgoing messages, where $q^i_{f\to V}$ is the numerator of (6.4).

Since we require uncertainty estimates on our predictions, we perform Bayesian linear regression from the random features to the output messages, which yields predictions close to those obtained by Gaussian process regression with the kernel in (6.9). The uncertainty estimate in this case will be the predictive variance. We assume prior and likelihood

$$\begin{aligned}\mathbf{w} &\sim \mathcal{N}\left(w; 0, I_{D_{\text{out}}}\sigma_0^2\right),\\ \mathsf{Y} \mid \mathsf{X}, \mathbf{w} &\sim \mathcal{N}\left(\mathsf{Y}; \mathbf{w}^\top\mathsf{X}, \sigma_y^2 I_N\right),\end{aligned}$$

where the output noise variance σ_y^2 captures the intrinsic stochasticity of the importance sampler used to generate Y. It follows that the posterior of w is given by [Bishop, 2006]

$$\begin{aligned}p(w|\mathsf{Y}) &= \mathcal{N}(w; \mu_w, \Sigma_w),\\ \Sigma_w &= \left(\mathsf{X}\mathsf{X}^\top\sigma_y^{-2} + \sigma_0^{-2} I\right)^{-1},\\ \mu_w &= \Sigma_w \mathsf{X}\mathsf{Y}^\top\sigma_y^{-2} = \left(XX^\top + \frac{\sigma_y^2}{\sigma_0^2} I\right)^{-1} XY^\top.\end{aligned}$$

The predictive distribution on the output y^* given an observation x^* is

$$\begin{aligned} p(\mathsf{y}^*|\mathsf{x}^*, \mathsf{Y}) &= \int p(\mathsf{y}^*|w, \mathsf{x}^*, \mathsf{Y}) p(w|\mathsf{Y})\, dw \\ &= \mathcal{N}\left(\mathsf{y}^*; \mathsf{x}^{*\top}\mu_w, \mathsf{x}^{*\top}\Sigma_w\mathsf{x}^* + \sigma_y^2\right). \end{aligned}$$

For simplicity, we treat each output (expected sufficient statistic) as a separate regression problem. Treating all outputs jointly can be achieved with a multi-output kernel [Álvarez et al., 2012].

Online Update We describe an online update for Σ_w and μ_w when observations (i.e., random features representing incoming messages) x_i arrive sequentially. We use $\cdot^{(N)}$ to denote a quantity constructed from N samples. The posterior covariance matrix at time $N+1$ is

$$\Sigma_w^{(N+1)} = \Sigma_w^{(N)} - \frac{\Sigma_w^{(N)}\mathsf{x}_{N+1}\mathsf{x}_{N+1}^\top\Sigma_w^{(N)}\sigma_y^{-2}}{1 + \mathsf{x}_{N+1}^\top\Sigma_w^{(N)}\mathsf{x}_{N+1}\sigma_y^{-2}},$$

meaning that it can be expressed as an inexpensive update of the covariance at time N. Updating Σ_w for all the D_y outputs costs $O((D_{\text{in}}D_{\text{out}} + D_{\text{out}}^2)D_y)$ per new observation. For $\mu_w = \Sigma_w\mathsf{X}\mathsf{Y}^\top\sigma_y^{-2}$, we maintain $\mathsf{X}\mathsf{Y}^\top \in \mathbb{R}^{D_{\text{out}}\times D_{\text{y}}}$, and update it at cost $O(D_{\text{in}}D_{\text{out}}D_y)$ as

$$\left(\mathsf{X}\mathsf{Y}^\top\right)^{(N+1)} = \left(\mathsf{X}\mathsf{Y}^\top + \mathsf{x}_{N+1}\mathsf{y}_{N+1}^\top\right).$$

Since we have D_y regression functions, for each tuple of incoming messages x^*, there are D_y predictive variances, $v_1^*, \ldots, v_{D_y}^*$, one for each output. Let $\{\tau_i\}_{i=1}^{D_y}$ be pre-specified predictive variance thresholds. Given a new input x^*, if $v_1^* > \tau_1$ or $\cdots$ or $v_{D_y}^* > \tau_{D_y}$ (the operator is uncertain), a query is made to the oracle to obtain a ground truth y^*. The pair $(\mathsf{x}^*, \mathsf{y}^*)$ is then used to update Σ_w and μ_w.

We refer to the new kernel-based message operator that learns just-in-time during inference as KJIT (kernel-based just-in-time learning).

6.4 Experiments

We evaluate our learned message operator using two different factors: the logistic factor, and the compound gamma factor. In the first and second experiments we demonstrate that the proposed message operator is capable of learning high-quality mappings from incoming to outgoing messages, and that the associated uncertainty estimates are reliable. The third and fourth experiments assess the performance of the operator as part of the full EP inference loop in two different models: approximating the logistic, and the compound gamma factors. Our final experiment demonstrates the ability of our learning process to reliably and quickly adapt to large shifts in the message distribution, as encountered during inference in a sequence of several real-world regression problems.

For all experiments we used Infer.NET [Minka et al., 2014] with its extensible factor interface for our own operator. We used the default settings of Infer.NET unless

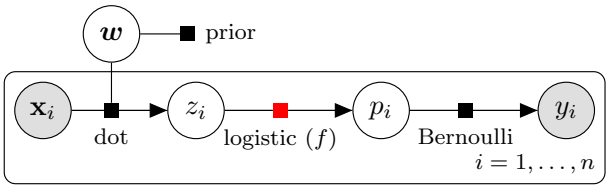


Figure 6.1: Factor graph for binary logistic regression. The kernel-based message operator learns to approximate the logistic factor highlighted in red. The two incoming messages are $m_{z_i \to f}(z_i) = \mathcal{N}(z_i; \mu, \sigma^2)$ and $m_{p_i \to f}(p_i) = \text{Beta}(p_i; \alpha, \beta)$. The prior distribution on the weight vector $\boldsymbol{w}$ is $\mathcal{N}(\mathbf{0}, I)$.

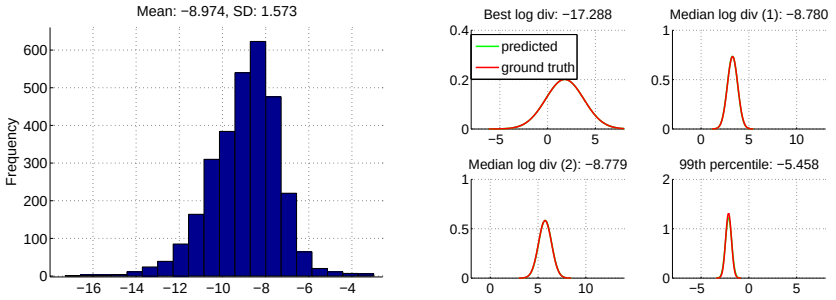
stated otherwise. The regression target is the marginal belief (numerator of (6.4)) in experiment 1,2,3 and 5. We set the regression target to the outgoing message in experiment 4. Given a marginal belief, the outgoing message can be calculated straightforwardly by dividing two messages (see (6.4)), both of which are in the same exponential family.

Experiment 1: Batch Learning As in [Heess et al., 2013, Eslami et al., 2014], we study the logistic factor $f(p|z) = \delta\left(p - \frac{1}{1+\exp(-z)}\right)$, where δ is the Dirac delta function, in the context of a binary logistic regression model (Figure 6.1). The factor is deterministic and there are two incoming messages: $m_{p_i \to f}(p_i) = \text{Beta}(p_i; \alpha, \beta)$ and $m_{z_i \to f}(z_i) = \mathcal{N}(z_i; \mu, \sigma^2)$, where $z_i = \boldsymbol{w}^\top \mathbf{x}_i$ represents the dot product between an observation $\mathbf{x}_i \in \mathbb{R}^d$ and the coefficient vector $\boldsymbol{w}$ whose posterior is to be inferred.

In this first experiment we simply learn a kernel-based operator to send the message $m_{f \to z_i}$. Following Eslami et al. [2014], we set d to 20, and generated 20 different datasets, each containing $\{(x_i, y_i)\}_{i=1}^n$ ($n = 300$) observations according to the model. For each dataset we ran EP, and collected incoming-outgoing message pairs in the first five iterations (i.e., the first five passes over all the nodes in the graph) of EP from Infer.NET's handcrafted implementation of the logistic factor. We partitioned the messages randomly into 5,000 training and 3,000 test messages, and learned a message operator to predict $m_{f \to z_i}$ as described in Section 6.3.3. Regularization and kernel parameters were chosen by leave-one-out cross validation. We set the number of random features to $D_{in} = 500$ and $D_{out} = 1,000$; empirically, we observed no significant improvements beyond 1,000 random features.

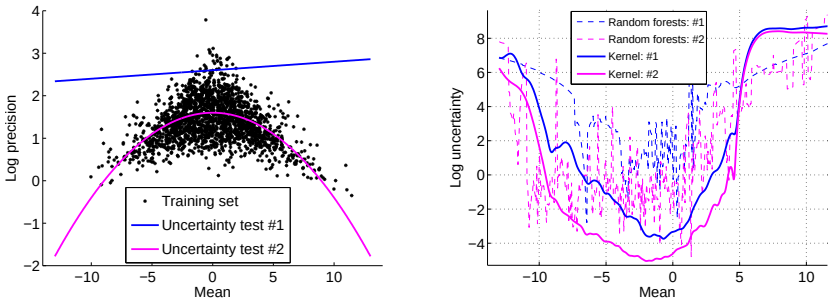
We report $\log \text{KL}[q_{f \to z_i} \| \hat{q}_{f \to z_i}]$ where $q_{f \to z_i}$ is the ground truth projected belief (numerator of (6.4)) and $\hat{q}_{f \to z_i}$ is the prediction. The histogram of the log KL errors is shown in Figure 6.2a; Figure 6.2b shows examples of predicted messages for different log KL errors. It is evident that the kernel-based operator does well in capturing the relationship between incoming and outgoing messages. The discrepancy with respect to the ground truth is barely visible even at the 99th percentile. See Section 6.7 for a comparison with other kernels and other methods.

Experiment 2: Uncertainty Estimates For the approximate message operator to perform well in a JIT learning setting, it is crucial to have reliable estimates of operator's predictive uncertainty in different parts of the space of incoming messages.



(a) log KL errors (b) Examples of predicted messages

Figure 6.2: (a) Log KL errors for predicting the projected beliefs to z_i. (b) Examples of predicted messages at different error levels.



(a) Parameters of $m_{z_i \to f}$ (b) Uncertainty estimates

Figure 6.3: (a) Incoming messages (each represented by a black dot) from z to f from 20 EP runs on the binary logistic regression problem, as shown in Figure 6.1. (b) Uncertainty estimates of the proposed kernel-based method (predictive variance) and Eslami et al.'s random forests (KL-based agreement of predictions of different trees) on the two uncertainty test sets shown. For testing, we fix the other incoming message $m_{p_i \to f}$ to $\text{Beta}(p_i; 1, 2)$.

To assess this property we compute the predictive variance using the same learned operator as used in Figure 6.2. The forward incoming messages $m_{z_i \to f}$ in the previously used training set are shown in Figure 6.3a. The backward incoming messages $m_{p_i \to f}$ are not displayed. Shown in the same plot are two curves (a blue line, and a pink parabola) representing two "uncertainty test sets": these are the sets of parameter pairs on which we wish to evaluate the uncertainty of the predictor, and pass through regions with both high and low densities of training samples. Figure 6.3b shows uncertainty estimates of our kernel-based operator and of random forests, where we fix $m_{p_i \to f}(p_i) := \text{Beta}(p_i; 1, 2)$ for testing. The implementation of the random forests closely follows Eslami et al. [2014].

From the figure, as the mean of the test message moves away from the region densely sampled by the training data, the predictive variance reported by the kernel method increases much more smoothly than that of the random forests. Further, our method clearly exhibits a higher uncertainty on the test set #1 than on the test set #2. This behaviour is desirable, as most of the points in test set #1 are either in a low density region or an unexplored region. These results suggest that the

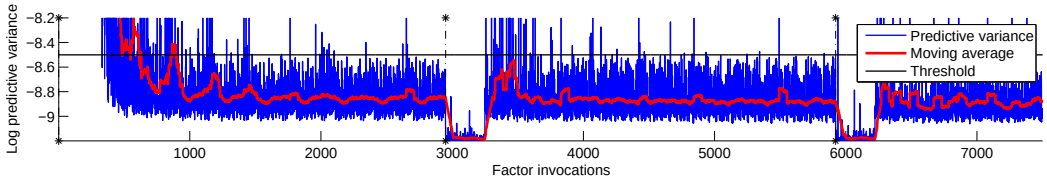


Figure 6.4: Uncertainty estimate of KJIT in its prediction of outgoing messages at each factor invocation, for the binary logistic regression problem. The black dashed lines indicate the start of a new inference problem.

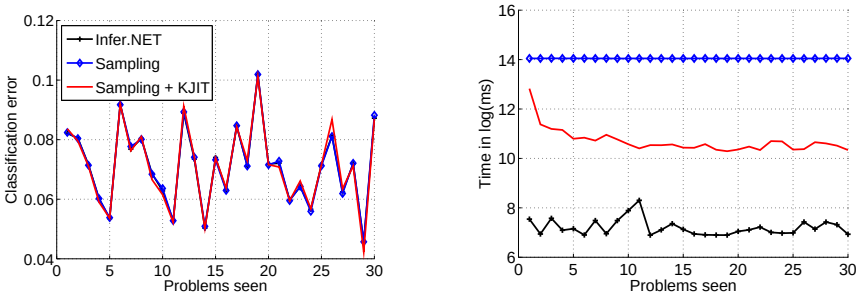
predictive variance is a robust criterion for querying the importance sampling oracle. One key observation is that the uncertainty estimates of the random forests are highly non-smooth; i.e., uncertainty on nearby points may vary wildly. As a result, a random forest-based JIT learner may still query the importance sampler oracle when presented with incoming messages similar to those in the training set, thereby wasting computation.

We have further checked that the predictive uncertainty of the regression function is a reliable indication of the error in KL divergence of the predicted outgoing messages. These results are given in Figure 6.7.1 of Section 6.7.

Experiment 3: Just-In-Time Learning In this experiment we test the message operator in the logistic regression model as part of the full EP inference loop in a just-in-time learning setting. We now learn two kernel-based message operators, one for each outgoing direction from the logistic factor. The data generation is the same as in the batch learning experiment. We sequentially presented the operator with 30 related problems, where a new set of observations $\{(x_i, y_i)\}_{i=1}^n$ was generated at the beginning of each problem from the model, while keeping $\boldsymbol{w}$ fixed. This scenario is common in practice: one is often given several sets of observations which share the same model parameter [Eslami et al., 2014]. As before, the inference target was $p(\boldsymbol{w}|\{(x_i, y_i)\}_{i=1}^n)$. We set the maximum number of EP iterations to 10 in each problem.

We employed a "mini-batch" learning approach in which the operator always consults the oracle in the first few hundred factor invocations for initial batch training. In principle, during the initial batch training, the operator can perform cross validation or type-II maximum likelihood estimation for parameter selection; however for computational simplicity we set the kernel parameters according to the median heuristic. Full detail of the heuristic is given in Section 6.5. The numbers of random features were $D_{\text{in}} = 300$ and $D_{\text{out}} = 500$. The output noise variance σ_y^2 was fixed to 10^{-4} and the uncertainty threshold on the log predictive variance was set to -8.5. To simulate a black-box setup, we used an importance sampler as the oracle rather than Infer.NET's factor implementation, where the proposal distribution was fixed to $\mathcal{N}(z; 0, 200)$ with 5×10^5 particles.

Figure 6.4 shows a trace of the predictive variance of KJIT in predicting the mean of each $m_{f \to z_i}$ upon each factor invocation. The black dashed lines indicate the start of a new inference problem. Since the first 300 factor invocations are for the initial



(a) Binary classification error (b) Inference time

Figure 6.5: Classification performance and inference times of all methods in the binary logistic regression problem.

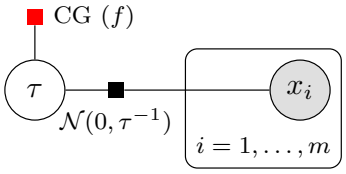
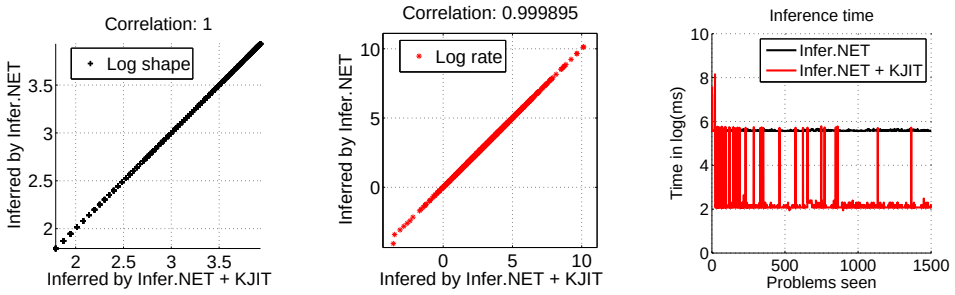


Figure 6.6: Factor graph of the compound gamma factor problem.

training, no uncertainty estimate is shown. From the trace, we observe that the uncertainty rapidly drops down to a stable point at roughly -8.8 and levels off after the operator sees about 1,000 incoming-outgoing message pairs, which is relatively low compared to approximately 3,000 message passings (i.e., 10 iterations $\times$ 300 observations) required for one problem. The uncertainty trace displays a periodic structure, repeating itself in every 300 factor invocations, corresponding to a full sweep over all 300 observations to collect incoming messages $m_{z_i \to f}$. The abrupt drop in uncertainty in the first EP iteration of each new problem is due to the fact that Infer.NET's inference engine initializes the message from $\boldsymbol{w}$ to have zero mean, leading to $m_{z_i \to f}$ also having a zero mean. Repeated encounters of such a zero mean incoming message reinforce the operator's confidence; hence the drop in uncertainty.

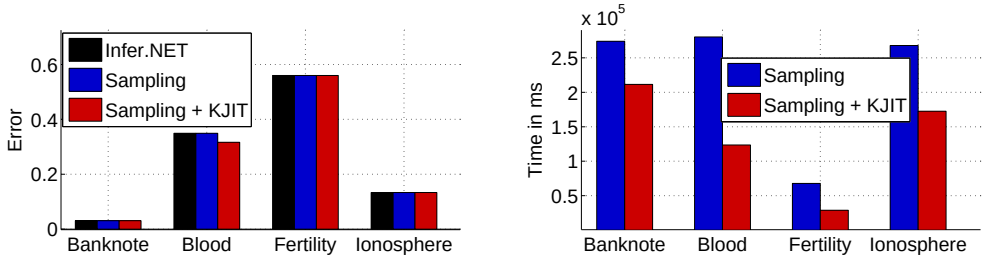
Figure 6.5a shows binary classification errors obtained by using the inferred posterior mean of $\boldsymbol{w}$ on a test set of size 10000 generated from the true underlying parameter. Included in the plot are the errors obtained by using only the importance sampler for inference ("Sampling"), and using the Infer.NET's hand-crafted logistic factor. The loss of KJIT matches well with that of the importance sampler and Infer.NET, suggesting that the inference accuracy is as good as these alternatives. Figure 6.5b shows the inference time required by all methods in each problem. While the inference quality is equally good, KJIT is orders of magnitude faster than the importance sampler.

Experiment 4: Compound Gamma Factor We next simulate the compound gamma factor, a heavy-tailed prior distribution on the precision of a Gaussian random variable. A variable τ is said to follow the compound gamma distribution if $\tau \sim \text{Gamma}(\tau; s_2, r_2)$ (shape-rate parameterization) and $r_2 \sim \text{Gamma}(r_2; s_1, r_1)$ where



(a) Inferred shape (b) Inferred rate (c) Inference time

Figure 6.7: Shape (a) and rate (b) parameters of the inferred posteriors in the compound gamma problem. (c) KJIT is able to infer equally good posterior parameters compared to Infer.NET, while requiring a runtime several orders of magnitude lower.



(a) Binary classification error (b) Inference time

Figure 6.8: Classification performance and inference times on the four UCI datasets.

(s_1, r_1, s_2) are parameters. The task we consider is to infer the posterior of the precision τ of a normally distributed variable $x \sim \mathcal{N}(x; 0, \tau^{-1})$ given realizations $\{x_i\}_{i=1}^m$ (see Figure 6.6 for the factor graph). We consider the setting $(s_1, r_1, s_2) = (1, 1, 1)$ which was used in Heess et al. [2013]. Infer.NET's implementation requires two gamma factors to specify the compound gamma. Here, we collapse them into one factor and let the operator learn to directly send an outgoing message $m_{f\to\tau}$ given $m_{\tau\to f}$, using Infer.NET as the oracle. The default implementation of Infer.NET relies on a quadrature method. As in Eslami et al. [2014], we sequentially presented a number of problems to our algorithm, where at the beginning of each problem, a random number of observations n from 10 to 100, and the parameter τ, were drawn from the model.

Figure 6.7a and Figure 6.7b summarize the inferred posterior parameters obtained from running only Infer.NET and Infer.NET + KJIT, i.e., KJIT with Infer.NET as the oracle. Figure 6.7c shows the inference time of both methods. The plots collectively show that KJIT can deliver posteriors in good agreement with those obtained from Infer.NET, at a much lower cost. Note that in this task only one message is passed to the factor in each problem. Figure 6.7c also indicates that KJIT requires fewer oracle consultations as more problems are seen.

Experiment 5: Classification Benchmarks In the final experiment, we demonstrate that our method for learning the message operator is able to detect changes in the

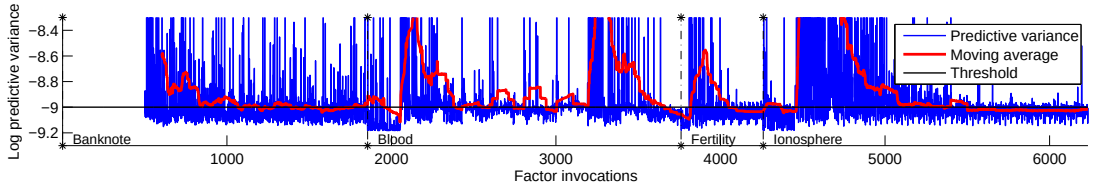


Figure 6.9: Uncertainty estimate of KJIT for outgoing messages on the four UCI datasets.

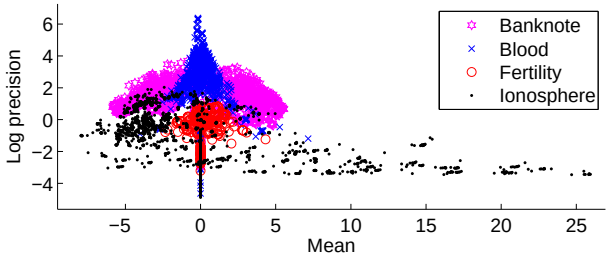


Figure 6.10: Distributions of incoming messages $m_{z_i \to f}$ to the logistic factor in four different UCI datasets (see Figure 6.1 for the factor graph).

distribution of incoming messages via its uncertainty estimate, and to subsequently update its prediction through additional oracle queries. The different distributions of incoming messages are achieved by presenting a sequence of different classification problems to our learner. We used four binary classification datasets from the UCI repository [Lichman, 2013]: banknote authentication, blood transfusion, fertility and ionosphere, in the same binary logistic regression setting as before. The operator was required to learn just-in-time to send outgoing messages $m_{f \to z_i}$ and $m_{f \to p_i}$ on the four problems presented in sequence. The training observations consisted of 200 data points subsampled from each dataset by stratified sampling. For the fertility dataset, which contains only 100 data points, we subsampled half the points. The remaining data were used as test sets. The uncertainty threshold was set to -9, and the minibatch size was 500. All other parameters were the same as in the earlier JIT learning experiment.

Classification errors on the test sets and inference times are shown in Figure 6.8a and Figure 6.8b, respectively. The results demonstrate that KJIT improves the inference time on all the problems without sacrificing inference accuracy. The predictive variance of each outgoing message is shown in Figure 6.9. An essential feature to notice is the rapid increase of the uncertainty after the first EP iteration of each problem. As shown in Figure 6.10, the distributions of incoming messages of the four problems are diverse. The sharp rise followed by a steady decrease of the uncertainty is a good indicator that the operator is able to promptly detect a change in input message distribution, and robustly adapt to this new distribution by querying the oracle.

Supplementary

6.5 Median Heuristic for the Gaussian Kernel on Mean Embeddings

In the proposed KJIT, there are two kernels: the inner kernel k for computing mean embeddings, and the outer Gaussian kernel κ defined on the mean embeddings. Both of the kernels depend on a number of parameters. In this section, we describe a heuristic to choose the kernel parameters. We emphasize that this heuristic is merely for computational convenience. A full parameter selection procedure like cross validation or evidence maximization will likely yield a better set of parameters. We use this heuristic in the initial mini-batch phase before the actual online learning.

Let $\{r_i^{(l)} \mid l = 1, \ldots, c, \text{ and } i = 1, \ldots, N\}$ be a set of N incoming message tuples collected during the mini-batch phase, from c variables neighboring the factor. Let $R_i := (r_i^{(l)})_{l=1}^c$ be the i^{th} tuple, and let $\mathsf{r}_i := \times_{l=1}^c r_i^{(l)}$ be the product of incoming messages in the i^{th} tuple. Define S_i and s_i to be the corresponding quantities of another tuple of messages. We will drop the subscript i when considering only one tuple.

Recall that the kernel on two tuples of messages R and S is given by

$$\begin{aligned}\kappa(R, S) = \kappa(\mathsf{r}, \mathsf{s}) &= \exp\left(-\frac{\|\mu_\mathsf{r} - \mu_\mathsf{s}\|_\mathcal{H}^2}{2\gamma^2}\right) \\ &= \exp\left(-\frac{1}{2\gamma^2}\langle\mu_\mathsf{r}, \mu_\mathsf{r}\rangle + \frac{1}{\gamma^2}\langle\mu_\mathsf{r}, \mu_\mathsf{s}\rangle - \frac{1}{2\gamma^2}\langle\mu_\mathsf{s}, \mu_\mathsf{s}\rangle\right),\end{aligned}$$

where $\langle\mu_\mathsf{r}, \mu_\mathsf{s}\rangle = \mathbb{E}_{x\sim\mathsf{r}}\mathbb{E}_{y\sim\mathsf{s}}k(\mathbf{x} - \mathbf{y})$. The inner kernel k is a Gaussian kernel defined on the domain $\mathcal{X} := \mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(c)}$ where $\mathcal{X}^{(l)}$ denotes the domain of $r^{(l)}$. For simplicity, we assume that $\mathcal{X}^{(l)}$ is one-dimensional. The Gaussian kernel k takes the form

$$k(\mathbf{x} - \mathbf{y}) = \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{y})^\top \Sigma^{-1}(\mathbf{x} - \mathbf{y})\right) = \prod_{l=1}^c \exp\left(-\frac{(x_l - y_l)^2}{2\sigma_l^2}\right),$$

where $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_c^2)$. The heuristic for choosing $\sigma_1^2, \ldots, \sigma_c^2$ and γ is as follows.

1. Set $\sigma_l^2 := \frac{1}{N}\sum_{i=1}^N \mathbb{V}_{x_l\sim r_i^{(l)}}[x_l]$ where $\mathbb{V}_{x_l\sim r_i^{(l)}}[x_l]$ denotes the variance of $r_i^{(l)}$.
2. With $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_c^2)$ as defined in the previous step, set

$$\gamma^2 := \text{median}\left(\{\|\mu_{\mathsf{r}_i} - \mu_{\mathsf{s}_j}\|^2\}_{i,j=1}^N\right).$$

6.6 Kernels and Random Features

This section gives details on other kernels of distributions that we explored, and describes their random feature representations.

6.6.1 Random Features

This section contains a summary of Rahimi and Recht [2007]'s random Fourier features for a translation invariant kernel.

A kernel $k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$ in general may correspond to an inner product in an infinite-dimensional space whose feature map ϕ cannot be explicitly computed. In Rahimi and Recht [2007], methods of computing an approximate feature map $\hat{\phi}$ were proposed. The approximate feature map is such that $k(\mathbf{x}, \mathbf{y}) \approx \hat{\phi}(\mathbf{x})^\top \hat{\phi}(\mathbf{y})$ (with equality in expectation) where $\hat{\phi}(\mathbf{x}), \hat{\phi}(\mathbf{y}) \in \mathbb{R}^D$ and D is the number of random features. High D yields a better approximation with higher computational cost. Assume $k(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y})$ (translation invariant) and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Random Fourier features $\hat{\phi}(\mathbf{x}) \in \mathbb{R}^D$ such that $k(\mathbf{x}, \mathbf{y}) \approx \hat{\phi}(\mathbf{x})^\top \hat{\phi}(\mathbf{y})$ are generated as follows:

1. Compute the Fourier transform $\hat{k}$ of the kernel k: $\hat{k}(\omega) = \frac{1}{2\pi} \int e^{-i\omega^\top \delta} k(\delta)\, d\delta$.
2. Draw D i.i.d. samples $\omega_1, \ldots, \omega_D \in \mathbb{R}^d$ from $\hat{k}$.
3. Draw D i.i.d samples $b_1, \ldots, b_D \in \mathbb{R}$ from $U[0, 2\pi]$ (uniform distribution).
4. $\hat{\phi}(\mathbf{x}) = \sqrt{\frac{2}{D}} \left(\cos\left(\omega_1^\top \mathbf{x} + b_1\right), \ldots, \cos\left(\omega_D^\top \mathbf{x} + b_D\right)\right)^\top \in \mathbb{R}^D$

This procedure is justified by the Bochner's theorem.

Theorem 6.1 (Bochner's theorem [Rudin, 2013])**.** *A continuous kernel $k(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y})$ on $\mathbb{R}^d$ is positive definite iff $k(\delta)$ is the Fourier transform of a non-negative measure.*

Bochner's theorem guarantees that the Fourier transform $\hat{k}$ can be seen as an unnormalized probability distribution. From this fact, we have

$$k(\mathbf{x} - \mathbf{y}) = \int \hat{k}(\omega) e^{i\omega^\top (\mathbf{x} - \mathbf{y})}\, d\omega = \mathbb{E}_{\omega \sim \hat{k}} \left[\eta_\omega(\mathbf{x}) \eta_\omega(\mathbf{y})^*\right],$$

where $i = \sqrt{-1}$, $\eta_\omega(\mathbf{x}) = e^{i\omega^\top \mathbf{x}}$ and $\cdot^*$ denotes the complex conjugate. Since both $\hat{k}$ and k are real, the complex exponential contains only the cosine terms. Drawing D samples lowers the variance of the approximation.

6.6.2 MV (Mean-Variance) Kernel

Assume there are c incoming messages $R := \left(r^{(l)}\right)_{l=1}^c$ and $S := \left(s^{(l)}\right)_{l=1}^c$. Assume that $\mathbb{E}_{r^{(l)}}[x] = m_l, \mathbb{V}_{r^{(l)}}[x] = v_l, \mathbb{E}_{s^{(l)}}[y] = \mu_l$, and $\mathbb{V}_{s^{(l)}}[y] = \sigma_l^2$. Incoming messages are not necessarily Gaussian. The MV (mean-variance) kernel is defined as a product kernel on means and variances.

$$\kappa_{\text{mv}}(R, S)$$

where k is a Gaussian kernel with unit width. The kernel κ_{mv} has $P := (w_1^m, \ldots, w_c^m, w_1^v, \ldots, w_c^v)$ as its parameters. With this kernel, we treat messages as finite dimensional vectors of their means and variances. Incoming messages $(s^{(i)})_{i=1}^c$ are represented as $(\mu_1, \ldots, \mu_c, \sigma_1^2, \ldots, \sigma_c^2)^\top$. This treatment reduces the problem of having distributions as inputs to the familiar problem of having input points from a Euclidean space. The random features of Rahimi and Recht [2007] can be applied straightforwardly.

6.6.3 Expected Product Kernel

Given two distributions $r(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{m}_r, V_r)$ and $s(\mathbf{y}) = \mathcal{N}(\mathbf{y}; \mathbf{m}_s, V_s)$ (d-dimensional Gaussian), the expected product kernel κ_{pro} is defined as

$$\kappa_{\text{pro}}(r, s) = \langle \mu_r, \mu_s \rangle_{\mathcal{H}} = \mathbb{E}_{\mathbf{x}\sim r} \mathbb{E}_{\mathbf{y}\sim s} k(\mathbf{x} - \mathbf{y}), \tag{6.12}$$

where $\mu_r := \mathbb{E}_{\mathbf{x}\sim r} k(\mathbf{x}, \cdot)$ is the mean embedding of r, and we assume that the kernel k associated with $\mathcal{H}$ is translation invariant i.e., $k(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y})$. The goal here is to derive random Fourier features for the expected product kernel. That is, we aim to find $\hat{\phi}$ such that $\kappa_{\text{pro}}(r, s) \approx \hat{\phi}(r)^\top \hat{\phi}(s)$ and $\hat{\phi}(\mathbf{x}) \in \mathbb{R}^D$. We first give some results which will be used to derive the Fourier features for inner product of mean embeddings.

Lemma 6.2. *If* $b \sim \mathcal{N}(b; 0, \sigma^2)$, *then* $\mathbb{E}[\cos(b)] = \exp\left(-\frac{1}{2}\sigma^2\right)$.

Proof. Consider the characteristic function of $x \sim \mathcal{N}(x; \mu, \sigma^2)$ which is given by

$$c_x(t) = \mathbb{E}_x \left[\exp\left(itx\right)\right] = \exp\left(it\mu - \frac{1}{2}\sigma^2 t^2\right).$$

For $\mu = 0, t = 1$, we have

$$c_b(1) = \mathbb{E}_{b\sim\mathcal{N}(0,\sigma^2)} \left[\exp(ib)\right] = \exp\left(-\frac{1}{2}\sigma^2\right) = \mathbb{E}_{b\sim\mathcal{N}(0,\sigma^2)} \left[\cos(b)\right],$$

where the imaginary part vanishes. □

We are ready to derive random features for the expected product kernel. From Rahimi and Recht [2007] which provides random features for $k(\mathbf{x} - \mathbf{y})$, we immediately have

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim r} \mathbb{E}_{\mathbf{y}\sim s} k(\mathbf{x} - \mathbf{y}) &\approx \mathbb{E}_{\mathbf{x}\sim r} \mathbb{E}_{\mathbf{y}\sim s} \frac{2}{D} \sum_{i=1}^{D} \cos\left(\mathbf{w}_i^\top \mathbf{x} + b_i\right) \cos\left(\mathbf{w}_i^\top \mathbf{y} + b_i\right) \\
&= \frac{2}{D} \sum_{i=1}^{D} \mathbb{E}_{\mathbf{x}\sim r} \cos\left(\mathbf{w}_i^\top \mathbf{x} + b_i\right) \mathbb{E}_{\mathbf{y}\sim s} \cos\left(\mathbf{w}_i^\top \mathbf{y} + b_i\right),
\end{aligned} \tag{6.13}$$

where $\{\mathbf{w}_i\}_{i=1}^D \sim \hat{k}(\mathbf{w})$ (Fourier transform of k) and $\{b_i\}_{i=1}^D \sim U[0, 2\pi]$.

Consider $\mathbb{E}_{\mathbf{x}\sim r} \cos\left(\mathbf{w}_i^\top \mathbf{x} + b_i\right)$, an expectation over $\mathbf{x}$ where b_i and $\mathbf{w}_i$ are given. Define $z_i = \mathbf{w}_i^\top \mathbf{x} + b_i$ so that $z_i \sim \mathcal{N}(z_i; \mathbf{w}_i^\top \mathbf{m}_r + b_i, \mathbf{w}_i^\top V_r \mathbf{w}_i)$. Let $d_i \sim \mathcal{N}(0, \mathbf{w}_i^\top V_r \mathbf{w}_i)$.

Then,

$$\begin{aligned}\mathbb{E}_{\mathbf{x}\sim r}\cos\left(\mathbf{w}_i^\top\mathbf{x}+b_i\right) &= \mathbb{E}_{z_i}[\cos(z_i)]\\ &= \mathbb{E}_{d_i}\left[\cos\left(d_i+\mathbf{w}_i^\top\mathbf{m}_r+b_i\right)\right]\\ &\overset{(a)}{=} \mathbb{E}_{d_i}\left[\cos(d_i)\cos(\mathbf{w}_i^\top\mathbf{m}_r+b_i)\right]-\mathbb{E}_{d_i}\left[\sin(d_i)\sin(\mathbf{w}_i^\top\mathbf{m}_r+b_i)\right]\\ &\overset{(b)}{=} \cos(\mathbf{w}_i^\top\mathbf{m}_r+b_i)\mathbb{E}_{d_i}\cos(d_i)\\ &\overset{(c)}{=} \cos(\mathbf{w}_i^\top\mathbf{m}_r+b_i)\exp\left(-\frac{1}{2}\mathbf{w}_i^\top V_r\mathbf{w}_i\right),\end{aligned} \tag{6.14}$$

where at (a) we use $\cos(\alpha+\beta)=\cos(\alpha)\cos(\beta)-\sin(\alpha)\sin(\beta)$. We have (b) because sine is an odd function and $\mathbb{E}_{d_i}\sin(d_i)=0$. The last equality (c) follows from Lemma 6.2. It follows from (6.13) and (6.14) that the random features $\hat{\phi}(r)\in\mathbb{R}^D$ are given by

$$\hat{\phi}(r)=\sqrt{\frac{2}{D}}\begin{pmatrix}\cos(\mathbf{w}_1^\top\mathbf{m}_r+b_1)\exp\left(-\frac{1}{2}\mathbf{w}_1^\top V_r\mathbf{w}_1\right)\\ \vdots\\ \cos(\mathbf{w}_D^\top\mathbf{m}_r+b_D)\exp\left(-\frac{1}{2}\mathbf{w}_D^\top V_r\mathbf{w}_D\right)\end{pmatrix}.$$

For a distribution r which is not normal, we only need to be able to compute $\mathbb{E}_{\mathbf{x}\sim r}\cos\left(\mathbf{w}_i^\top\mathbf{x}+b_i\right)$. With $\hat{\phi}(r)$, we have $\kappa_{\text{pro}}(r,s)\approx\hat{\phi}(r)^\top\hat{\phi}(s)$ with equality in expectation.

Closed-Form Expression for Gaussian Case For reference, if $r(\mathbf{x})=\mathcal{N}(\mathbf{x};\mathbf{m}_r,V_r)$ and $s(\mathbf{y})=\mathcal{N}(\mathbf{y};\mathbf{m}_s,V_s)$ and k is a Gaussian kernel, (6.12) can be computed in closed-form. Assume $k(\mathbf{x}-\mathbf{y})=\exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{y})^\top\Sigma^{-1}(\mathbf{x}-\mathbf{y})\right)$ where the positive definite matrix Σ is the kernel parameter. Then,

$$\kappa_{\text{pro}}(r,s)=\mathbb{E}_{\mathbf{x}\sim r}\mathbb{E}_{\mathbf{y}\sim s}k(\mathbf{x}-\mathbf{y})=\sqrt{\frac{\det(D_{rs})}{\det(\Sigma^{-1})}}\exp\left(-\frac{1}{2}(\mathbf{m}_r-\mathbf{m}_s)^\top D_{rs}(\mathbf{m}_r-\mathbf{m}_s)\right),$$
$$D_{rs}:=(V_r+V_s+\Sigma)^{-1},$$

where $\det(A)$ denotes the determinant of A.

6.6.4 Product and Sum Kernels on Mean Embeddings

Previously, we have defined an expected product kernel on single distributions. One way to define a kernel between two tuples of more than one incoming message is to take a product of the kernels defined on each message.

Let $\mu_{r^{(l)}}:=\mathbb{E}_{r^{(l)}(a)}k^{(l)}(\cdot,a)$ be the mean embedding [Smola et al., 2007] of the distribution $r^{(l)}$ into RKHS $\mathcal{H}^{(l)}$ induced by the kernel k. Assume $k^{(l)}=k^{(l)}_{\text{gauss}}$ (Gaussian kernel) and assume there are c incoming messages $R:=(r^{(i)}(a^{(i)}))_{i=1}^c$ and $S:=(s^{(i)}(b^{(i)}))_{i=1}^c$. A product of expected product kernels is defined as

$$\kappa^{\times}_{\text{pro}}(R,S):=\left\langle\bigotimes_{l=1}^{c}\mu_{r^{(l)}},\bigotimes_{l=1}^{c}\mu_{s^{(l)}}\right\rangle_{\otimes_l\mathcal{H}^{(l)}}$$

$$= \prod_{l=1}^{c} \mathbb{E}_{r^{(l)}(a^{(l)})} \mathbb{E}_{s^{(l)}(b^{(l)})} k^{(l)}_{\text{gauss}} \left(a^{(l)}, b^{(l)} \right) \approx \hat{\phi}(R)^\top \hat{\phi}(S),$$

where $\hat{\phi}(R)^\top \hat{\phi}(S) = \prod_{l=1}^{c} \hat{\phi}^{(l)}(r^{(l)})^\top \hat{\phi}^{(l)}(s^{(l)})$. The feature map $\hat{\phi}^{(l)}(r^{(l)})$ can be estimated by applying the random Fourier features to $k^{(l)}_{\text{gauss}}$ and taking the expectations $\mathbb{E}_{r^{(l)}(a)} \mathbb{E}_{s^{(l)}(b)}$. The final feature map is $\hat{\phi}(R) = \hat{\phi}^{(1)}(r^{(1)}) \circledast \hat{\phi}^{(2)}(r^{(2)}) \circledast \cdots \circledast \hat{\phi}^{(c)}(r^{(c)}) \in \mathbb{R}^{D^c}$, where $\circledast$ denotes a Kronecker product and we assume that $\hat{\phi}^{(l)} \in \mathbb{R}^D$ for $l \in \{1, \ldots, c\}$.

If we define the kernel as the sum of c kernels instead, we have

$$\kappa^{+}_{\text{pro}}(R, S) = \sum_{l=1}^{c} \langle \mu_{r^{(l)}}, \mu_{s^{(l)}} \rangle_{\mathcal{H}^{(l)}} \approx \sum_{l=1}^{c} \hat{\phi}^{(l)}(r^{(l)})^\top \hat{\phi}^{(l)}(s^{(l)}) = \hat{\varphi}(R)^\top \hat{\varphi}(S)$$

where $\hat{\varphi}(R) := \left(\hat{\phi}^{(1)}(r^{(1)})^\top, \ldots, \hat{\phi}^{(c)}(r^{(c)})^\top \right)^\top \in \mathbb{R}^{cD}$.

6.7 More Details on Experiment 1: Batch Learning

There are a number of kernels on distributions we may use for just-in-time learning. To find the most suitable kernel, we compare the performance of each on a collection of incoming and output messages at the logistic factor in the binary logistic regression problem i.e., same problem as in experiment 1 in the main text. All messages are collected by running EP 20 times on generated toy data. Only messages in the first five EP iterations are considered, since messages passed in the early phase of EP vary more than in a near-convergence phase. The regression output to be learned is the numerator of (6.4).

A training set of 5000 messages and a test set of 3000 messages are obtained by subsampling all the collected messages. Where random features are used, kernel widths and regularization parameters are chosen by leave-one-out cross validation. To get a good sense of the approximation error from the random features, we also compare with incomplete Cholesky factorization (denoted by IChol), a widely used Gram matrix approximation technique. We use hold-out cross validation with randomly chosen training and validation sets for parameter selection, and kernel ridge regression in its dual form when the incomplete Cholesky factorization is used.

Let f be the logistic factor and $m_{f \to i}$ be an outgoing message. Let $q_{f \to i}$ be the ground truth belief message (numerator) associated with $m_{f \to i}$. Following Heess et al. [2013], Eslami et al. [2014], the error metric we use is $\log \mathrm{KL}[q_{f \to i} \,||\, \hat{q}_{f \to i}]$ where $\hat{q}_{f \to i}$ is the belief message estimated by a learned message operator. Table 6.7.1 reports the means and standard deviations of the log KL-divergence.

The MV kernel is defined in Section 6.6.2. Here, product (sum) of expected product kernels refers to a product (sum) of kernels, where each is an expected product kernel defined on one incoming message (see Sections 6.6.3, and 6.6.4). Evidently, the

Table 6.7.1: Means and standard deviations of the log KL-divergence between the ground truth and predicted outgoing messages using different message operators. RFFs stands for random Fourier features.

	Mean log KL	**SD of log KL**
RFFs + MV Kernel	-6.96	1.67
RFFs + Expected product kernel on joint embeddings	-2.78	1.82
RFFs + Sum of expected product kernels	-1.05	1.93
RFFs + Product of expected product kernels	-2.64	1.65
RFFs + Gaussian kernel on joint embeddings (KJIT)	**-8.97**	1.57
IChol + sum of Gaussian kernel on embeddings	-2.75	2.84
IChol + Gaussian kernel on joint embeddings	-8.71	1.69
Breiman's random forests [Breiman, 2001]	-8.69	1.79
Extremely randomized trees [Geurts et al., 2006]	-8.90	1.59
Eslami et al. [2014]'s random forests	-6.94	3.88

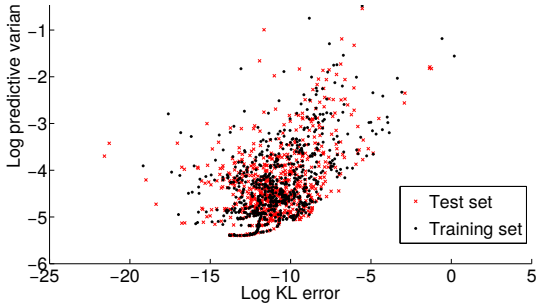


Figure 6.7.1: KL-divergence error versus predictive variance for predicting the mean of $m_{f \to z_i}$ (normal distribution) in the logistic factor problem.

Gaussian kernel on joint mean embeddings (see (6.9)) performs significantly better than other kernels. Besides the proposed method, we also compare the message prediction performance to Breiman's random forests [Breiman, 2001], extremely randomized trees [Geurts et al., 2006], and Eslami et al. [2014]'s random forests. We use scikit-learn toolbox [Pedregosa et al., 2011] for the extremely randomized trees and Breiman's random forests. For Eslami et al. [2014]'s random forests, we reimplemented the method as closely as possible according to the description given in Eslami et al. [2014]. In all cases, the number of trees is set to 64. Empirically we observe that decreasing the log KL error below -8 will not yield a noticeable performance gain in the actual EP.

To verify that the uncertainty estimates given by KJIT coincide with the actual predictive performance (i.e., accurate prediction when confident), we plot the predictive variance against the log KL-divergence error on both the training and test sets. The results are shown in Figure 6.7.1. The uncertainty estimates show a positive correlation with the KL-divergence errors. It is instructive to note that no point lies at the bottom right i.e., making a large error while being confident. The fact that the errors on the training set are roughly the same as the errors on the test set indicates that the operator does not overfit.

Chapter 7

Conclusions and Future Work

In this thesis, we consider learning explicit features for statistical tests and Bayesian inference with expectation propagation. We consider three hypothesis testing problems: two-sample testing, independence testing, and goodness-of-fit testing. In the setting of two-sample testing (Chapter 3), we have considered two tests: the mean embedding (ME) test, and the smooth characteristic function (SCF) test. In the ME test, the features correspond to the spatial locations at which the witness function (i.e., difference of mean embeddings of the two distributions) is evaluated. In the SCF test, the features are in the frequency domain and are where the difference of smoothed characteristic functions is evaluated. The features in these two tests thus indicate where the two underlying distributions differ in the spatial and frequency domains, respectively. In the independence test (Chapter 4), the features are points in the joint domain of the two random variables X and Y, indicating the regions in which the joint distribution and the product of marginal distributions of X and Y differ most. In the goodness-of-fit test (Chapter 5), the features are points in the domain of the model density, indicating where the model does not fit the observed data. In all these tests, features can be automatically optimized so as to maximize (a lower bound on) the rate of detecting the differences of two distributions. All resulting tests, and their associated optimization procedures have runtime complexity which is linear in the sample size. We have shown that the tests are consistent for any number of features. We observe empirically that all proposed linear-time tests have high performance, comparable to, or in some case exceeding that of their respective competing quadratic-time tests.

In the second part on automated expectation propagation (EP) (Chapter 6), we have proposed a method for learning a message operator mapping from a tuple of incoming EP messages to an outgoing message. A learned message operator can be used in place of computationally demanding Monte Carlo estimates of outgoing messages. Our operator has two main advantages: it can reliably evaluate the uncertainty of its prediction, so that it only consults a more expensive oracle when it is uncertain, and it can efficiently update its mapping online, so that it learns from these additional consultations. Once trained, the learned mapping performs as well as the oracle

mapping, but at a far lower computational cost. This is in large part due to a novel two-stage random feature representation of the input messages.

Future Work

For automated EP, one possible topic of research is hyperparameter selection. At present, hyperparameters are learned on an initial mini-batch of data, however a better option would be to adapt them online as more data are seen. There are several possible future directions for the three proposed linear-time tests.

Bahadur Slope as the Optimization Objective The optimization objective for the features in our proposed goodness-of-fit test (FSSD) is based on its test power (see (5.5)). To recall, the objective is $\mathrm{FSSD}^2/\sigma_{H_1}$ where FSSD is the population test statistic which depends on the features and the kernel, and σ_{H_1} is the standard deviation of the asymptotic normal distribution of $\sqrt{n}\left(\widehat{\mathrm{FSSD}^2} - \mathrm{FSSD}^2\right)$ assuming H_1 is true (see Proposition 5.3). The approximate Bahadur slope of $n\widehat{\mathrm{FSSD}^2}$ is $c^{(\mathrm{FSSD})} := \mathrm{FSSD}^2/\omega_1$, where ω_1 is the maximum eigenvalue of the covariance matrix under H_0 (see Theorem 5.10), which depends on the kernel and the features. It would be interesting to optimize features by maximizing $c^{(\mathrm{FSSD})}$, and compare the obtained features and the features learned by the procedure proposed in Section 5.3.2. Further, approximate Bahadur slopes for the proposed two-sample and independence test statistics may also be derived. A study on features obtained by optimizing their respective approximate Bahadur slopes will be an interesting topic of research.

Linear-Time Test for Model Comparison The proposed goodness-of-fit test in Chapter 5 determines whether an observed sample $\{\mathbf{x}_i\}_{i=1}^n$ follows a known density p (model). In some cases, especially with real data, the hypothetical models are known a priori to be wrong. The relevant question then becomes: given a sample $\{\mathbf{x}_i\}_{i=1}^n$, and two density functions p_1 and p_2 representing two competing models known to be wrong, which one fits the data better? A similar question was addressed before by Bounliphone et al. [2015] where the test relies on samples $\{\mathbf{y}_i\}_{i=1}^n \sim p_1$ and $\{\mathbf{z}_i\}_{i=1}^n \sim p_2$ from the two models. In our case, the goal is to design a linear-time test for model comparison, which directly makes use of the two known density functions. Following Bounliphone et al. [2015], a potential approach is to propose the null hypothesis $H_0 : \mathrm{FSSD}_1^2 \leq \mathrm{FSSD}_2^2$ (i.e., p_1 fits better) against the alternative hypothesis $H_1 : \mathrm{FSSD}_1^2 > \mathrm{FSSD}_2^2$, where FSSD_i^2 denotes the discrepancy between the sample $\{\mathbf{x}_i\}_{i=1}^n$ and model p_i, as measured by FSSD^2 (defined in Theorem 5.2). It follows that the empirical test statistic is $S_n := \widehat{\mathrm{FSSD}_1^2} - \widehat{\mathrm{FSSD}_2^2}$, where $\widehat{\mathrm{FSSD}_1^2}$ and $\widehat{\mathrm{FSSD}_2^2}$ are estimates based on the sample. The asymptotic null distribution can be derived by considering the joint distribution of $\widehat{\mathrm{FSSD}_1^2}$ and $\widehat{\mathrm{FSSD}_2^2}$ i.e., they are dependent as the same sample is used to estimate them. It will be interesting to design the procedure so that the learned features indicate where (in the data domain) p_2 fits the data better than p_1.

Appendix A

Appendix

This chapter summarizes well-known results that we use in this thesis.

A.1 U-Statistics

U-statistics, originally proposed in Hoeffding [1948], form a general class of unbiased estimators. Many commonly used estimators can be written as a U-statistic including mean, covariance, as well as the MMD's unbiased estimator in (2.11). In this section, we define U-statistics, and provide some known results regarding the convergence properties, and asymptotic distributions under different conditions. Good references on U-statistics include Hoeffding [1948], Kowalski and Tu [2008], Serfling [2009].

Definition A.1 (U-statistic (Hoeffding [1948], Serfling [2009, Section 5.1]))**.** Consider a symmetric function $h(\mathbf{a}_1, \ldots, \mathbf{a}_m)$ of m arguments i.e., $h(\mathbf{a}_1, \ldots, \mathbf{a}_m)$ is invariant to the permutation of the m arguments. Let $\{\mathbf{x}_i\}_{i=1}^n$ be an i.i.d. sample from a multivariate distribution F, where $n \geq m$. A one-sample order-m U-statistic is the statistic of the form

$$U_n := \binom{n}{m}^{-1} \sum_{\mathbf{c} \in C_m^n} h(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_m}), \tag{A.1}$$

where $C_m^n := \{(i_1, \ldots, i_m) \mid 1 \leq i_1 < \cdots < i_m \leq n\}$ so that $\sum_{\mathbf{c} \in C_m^n}$ denotes summation over the $\binom{n}{m}$ combinations of m distinct elements $\{i_1, \ldots, i_m\} \subset \{1, \ldots, n\}$. The statistic is an unbiased estimate of

$$\begin{aligned} \theta(F) &= \int \cdots \int h(\mathbf{x}_1, \ldots, \mathbf{x}_m) \, \mathrm{d}F(\mathbf{x}_1) \cdots \mathrm{d}F(\mathbf{x}_m) \\ &= \mathbb{E}_{\mathbf{x}_1 \sim F} \cdots \mathbb{E}_{\mathbf{x}_m \sim F} [h(\mathbf{x}_1, \ldots, \mathbf{x}_m)] \end{aligned} \tag{A.2}$$

The function h is known as the *U-statistic kernel* of θ. Where there is ambiguity, we will reserve the term kernel for a positive definite kernel, and refer to a U-statistic kernel as a U-statistic *core*.

Notations For brevity, we will write the expectations $\mathbb{E}_{\mathbf{x}_1 \sim F} \cdots \mathbb{E}_{\mathbf{x}_m \sim F}$ in (A.2) as $\mathbb{E}_F$. We write θ for $\theta(F)$. Similarly, the variance over all the m variables is denoted by $\mathbb{V}_F$. Define $h_c(\mathbf{x}_1, \ldots \mathbf{x}_c) = \mathbb{E}_F[h(\mathbf{x}_1, \ldots, \mathbf{x}_m) \mid \mathbf{x}_1, \ldots, \mathbf{x}_c]$ for $c \in \{1, \ldots, m\}$ so that $h_m = h$.

Define $\zeta_0 := 0$ and, for $c \in \{1, \dots m\}$, let $\zeta_c := \mathbb{V}_F[h_c(\mathbf{x}_1, \dots, \mathbf{x}_c)]$. Let $\tilde{h}_c := h_c - \theta$ for $c \in \{1, \dots, m\}$ so that $\tilde{h}_m = \tilde{h} = h - \theta$. Convergence in distribution is denoted by $\xrightarrow{d}$.

Lemma A.2 (Finite-sample variance of U-statistics [Serfling, 2009, Lemma A, p. 183])**.** *The variance of U_n is given by*

$$\mathbb{V}_F[U_n] = \binom{n}{m}^{-1} \sum_{c=1}^{m} \binom{m}{c} \binom{n-m}{m-c} \zeta_c, \tag{A.3}$$

and satisfies

1. $\frac{m^2}{n}\zeta_1 \leq \mathbb{V}_F[U_n] \leq \frac{m}{n}\zeta_m$,
2. $(n+1)\,\mathbb{V}_F[U_{n+1}] \leq n\mathbb{V}_F[U_n]$,
3. $\mathbb{V}_F[U_n] = \frac{m^2\zeta_1}{n} + \mathcal{O}(n^{-2})$, $n \to \infty$.

The variance expression in (A.3) holds for any finite sample size n, whereas item 3 in Lemma A.2 gives the asymptotic variance. The latter is useful for deriving the asymptotic distributions of many test statistics. It can be shown that $0 = \zeta_0 \leq \zeta_1 \leq \cdots \leq \zeta_m = \mathbb{V}_F[h(\mathbf{x}_1, \dots, \mathbf{x}_m)] < \infty$ [Serfling, 2009, Section 5.2.1].

There are two cases to consider in deriving the asymptotic distributions of U_n as $n \to \infty$: when $\zeta_1 > 0$ and, when $\zeta_1 = 0$. When $\zeta_1 = 0$, the U-statistic U_n is said to be *degenerate*. We start with the asymptotic distribution of a degenerate U-statistic.

Lemma A.3 (Asymptotic distribution of degenerate U-statistics [Serfling, 2009, Section 5.5.2][1])**.** *Define an operator A acting on a function $g \in L_2(\mathbb{R}^d, F)$ by*

$$Ag(\mathbf{x}) = \int_{\mathbb{R}^d} \tilde{h}_2(\mathbf{x}, \mathbf{y}) g(\mathbf{y})\, \mathrm{d}F(\mathbf{y}),$$

where $\mathbf{x} \in \mathbb{R}^d$. Let $\lambda_1, \lambda_2, \dots$ be real eigenvalues of A satisfying $Ag(\mathbf{x}) = \lambda_i g(\mathbf{x})$. If $\mathbb{E}_F[h^2(\mathbf{x}_1, \dots, \mathbf{x}_m)] < \infty$ and $\zeta_1 = 0 < \zeta_2$, then

$$n(U_n - \theta) \xrightarrow{d} \frac{m(m-1)}{2} \sum_{i=1}^{\infty} \lambda_i (Z_i^2 - 1),$$

where $Z_1, Z_2, \dots \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$.

Lemma A.3 states that under appropriate conditions, a degenerate U-statistic asymptotically has the same distribution as an infinite weighted sum of chi-squared random variables (up to a constant shift by $\frac{m(m-1)}{2}\sum_{i=1}^{\infty}\lambda_i$). The asymptotic distribution of U_n when $\zeta_1 > 0$ is given in Lemma A.4.

Lemma A.4 (Asymptotic distribution of non-degenerate U-statistics [Serfling, 2009, Section 5.5.1])**.** *If $\mathbb{E}_F[h^2(\mathbf{x}_1, \dots, \mathbf{x}_m)] < \infty$ and $\zeta_1 > 0$, then $\sqrt{n}(U_n - \theta) \xrightarrow{d} \mathcal{N}(0, m^2\zeta_1)$.*

[1] The discussion in Serfling [2009, Section 5.5.2] only considers univariate random variables. However, the general theory of U-statistics holds for multivariate random variables [Hoeffding, 1948].

When the U-statistic core is bounded, a finite-sample bound can be derived and is given in Lemma A.5. This finite-sample bound generalizes Hoeffding' inequality for sum of independent random variables.

Lemma A.5 (A bound for U-statistics [Serfling, 2009, Theorem A, p. 201]). *Let $h(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ be a U-statistic kernel for an m-order U-statistic such that $h(\mathbf{x}_1, \ldots, \mathbf{x}_m) \in [a, b]$ where $a \leq b < \infty$. Let $U_n = \binom{n}{m}^{-1} \sum_{i_1 < \cdots < i_m} h(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_m})$ be a U-statistic computed with a sample of size n, where the summation is over the $\binom{n}{m}$ combinations of m distinct elements $\{i_1, \ldots, i_m\}$ from $\{1, \ldots, n\}$. Then, for $t > 0$ and $n \geq m$,*

$$\mathbb{P}(U_n - \mathbb{E}h(\mathbf{x}_1, \ldots, \mathbf{x}_m) \geq t) \leq \exp\left(-2\lfloor n/m \rfloor t^2/(b-a)^2\right),$$
$$\mathbb{P}(|U_n - \mathbb{E}h(\mathbf{x}_1, \ldots, \mathbf{x}_m)| \geq t) \leq 2\exp\left(-2\lfloor n/m \rfloor t^2/(b-a)^2\right),$$

where $\lfloor x \rfloor$ denotes the greatest integer which is smaller than or equal to x. Hoeffding's inequality is a special case when $m = 1$.

Multivariate U-Statistics

An extension to the U-statistics defined previously is multivariate U-statistics [Hoeffding, 1948, Kowalski and Tu, 2008], where U_n now has multiple outputs. Let

$$\mathbf{h}(\mathbf{a}_1, \ldots, \mathbf{a}_m) = \left(h^{(1)}(\mathbf{a}_1, \ldots, \mathbf{a}_m), \ldots, h^{(J)}(\mathbf{a}_1, \ldots, \mathbf{a}_m)\right) \in \mathbb{R}^J$$

be a stack of J U-statistic kernels.[2] Let $\{\mathbf{x}_i\}_{i=1}^n$ be an i.i.d. sample drawn from F. A one-sample order-m *multivariate U-statistic* is defined as

$$\mathbf{U}_n := \binom{n}{m}^{-1} \sum_{\mathbf{c} \in C_m^n} \mathbf{h}(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_m}) = \left(U_n^{(1)}, \ldots, U_n^{(J)}\right)^\top,$$

where $C_m^n := \{(i_1, \ldots, i_m) \mid 1 \leq i_1 < \cdots < i_m \leq n\}$. As in the univariate case in Definition A.1, $\mathbf{U}_n$ is an unbiased estimate of $\boldsymbol{\theta} = \left(\theta^{(1)}, \ldots, \theta^{(J)}\right)^\top := \mathbb{E}_F[\mathbf{U}_n] \in \mathbb{R}^J$. Similarly to the univariate case, we define $\mathbf{h}_c(\mathbf{x}_1, \ldots \mathbf{x}_c) := \mathbb{E}_F[\mathbf{h}(\mathbf{x}_1, \ldots, \mathbf{x}_m) \mid \mathbf{x}_1, \ldots, \mathbf{x}_c]$, and $\widetilde{\mathbf{h}}_c := \mathbf{h}_c - \boldsymbol{\theta}$ for $c \in \{1, \ldots, m\}$. Let $\zeta_0^{(j)} := 0$, and $\zeta_c^{(j)} := \mathbb{V}_F[h_c^{(j)}(\mathbf{x}_1, \ldots, \mathbf{x}_c)]$ for $c \in \{1, \ldots m\}$ and $j \in \{1, \ldots, J\}$. Further define

$$\begin{aligned}
\zeta_c^{(l,j)} &:= \mathrm{cov}_F[h_c^{(l)}(\mathbf{x}_1, \ldots, \mathbf{x}_c), h_c^{(j)}(\mathbf{x}_1, \ldots, \mathbf{x}_c)] \\
&= \mathbb{E}_F[\tilde{h}_c^{(l)}(\mathbf{x}_1, \ldots, \mathbf{x}_c)\tilde{h}_c^{(j)}(\mathbf{x}_1, \ldots, \mathbf{x}_c)] \\
&= \mathbb{E}_F[h_c^{(l)}(\mathbf{x}_1, \ldots, \mathbf{x}_c)h_c^{(j)}(\mathbf{x}_1, \ldots, \mathbf{x}_c)] - \theta^{(l)}\theta^{(j)}
\end{aligned}$$

for $l, j \in \{1, \ldots, J\}$ and $c \in \{1, \ldots, m\}$. Note that if $l = j$, then $\zeta_c^{(l,j)} = \zeta_c^{(l)}$.

Lemma A.6 (Asymptotic distribution of multivariate U-statistics (Hoeffding [1948, Theorem 7.1], Kowalski and Tu [2008, p. 255], Lehmann [1999, Theorem 6.1.5])). *Define*

[2] In its full generality, each of the J U-statistic kernels can be of different orders [Hoeffding, 1948]. We have omitted this case for brevity since it is not required in this study.

$\boldsymbol{\Sigma} \in \mathbb{R}^{J \times J}$ *such that* $\Sigma_{ij} := \zeta_1^{(i,j)}$. *If* $\boldsymbol{\Sigma}$ *is positive definite, then*

$$\sqrt{n}(\mathbf{U}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, m^2\boldsymbol{\Sigma}).$$

For the limiting Gaussian distribution to be non-degenerate, the $\boldsymbol{\Sigma}$ has to be positive definite. In particular, this condition requires that $0 < \zeta_1^{(j)} = \mathbb{V}_F[h_1^{(j)}(\mathbf{x}_1)] < \infty$ for all $j \in \{1, \ldots, J\}$ i.e., the j^{th} U-statistic as defined by the kernel $h^{(j)}$ is not degenerate.

Bibliography

M. A. Álvarez, L. Rosasco, and N. D. Lawrence. Kernels for vector-valued functions: A review. *Found. Trends Mach. Learn.*, 4(3):195–266, Mar. 2012. 127

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, 2003. 33, 70

M. A. Arcones and E. Gine. On the bootstrap of U and V statistics. *The Annals of Statistics*, pages 655–674, 1992. 24

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002. 64, 73

R. R. Bahadur. Stochastic comparison of tests. *The Annals of Mathematical Statistics*, 31(2): 276–295, 1960. 101, 102, 103, 104

L. Baringhaus and C. Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88:190–206, 2004. 30

L. Baringhaus and N. Henze. A consistent test for multivariate normality based on the empirical characteristic function. *Metrika*, 35:339–348, 1988. 93

S. Barthelmé and N. Chopin. ABC-EP: Expectation propagation for likelihood-free Bayesian computation. In *ICML*, pages 289–296, 2011. 14, 116, 120, 121

J. Beirlant, L. Györfi, and G. Lugosi. On the asymptotic normality of the l_1- and l_2-errors in histogram density estimation. *Canadian Journal of Statistics*, 22:309–318, 1994. 93

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004. 12, 17, 20

T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *International Conference on Music Information Retrieval (ISMIR)*, 2011. 76

R. Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013. 114

M. Bilodeau and D. Brenner. *Theory of multivariate statistics*. Springer Science & Business Media, 2008. 36

S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python*. O'Reilly Media, 1st edition, 2009. 47

C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. 117, 126

W. Bounliphone, E. Belilovsky, M. B. Blaschko, I. Antonoglou, and A. Gretton. A test of relative similarity for model selection in generative models. *ArXiv e-prints*, Nov. 2015. 142

O. Bousquet. New approaches to statistical learning theory. *Annals of the Institute of Statistical Mathematics*, 55(2):371–389, 2003. 56

A. Bowman and P. Foster. Adaptive smoothing and density based tests of multivariate normality. *Journal of the American Statistical Association*, 88:529–537, 1993. 93

L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, Oct. 2001. 140

C. Carmeli, E. De Vito, A. Toigo, and V. Umanità. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 08(01):19–61, Jan. 2010. 26, 94, 96

A. Christmann and I. Steinwart. Universal kernels on non-standard input spaces. In *NIPS*, pages 406–414, 2010. 25, 116, 123, 124

K. Chwialkowski and A. Gretton. A kernel independence test for random processes. In *ICML*, volume 32, pages 1422–1430, 2014. 11

K. Chwialkowski, D. Sejdinovic, and A. Gretton. A wild bootstrap for degenerate kernel tests. In *NIPS*, pages 3608–3616, 2014. 11, 99

K. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representations of probability measures. In *NIPS*, pages 1972–1980, 2015. 12, 27, 29, 30, 31, 32, 33, 34, 35, 39, 40, 41, 65, 67, 70, 94

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *ICML*, pages 2606–2615, 2016. 12, 94, 95, 96, 97, 98, 108

G. P. Cinzia Carota and N. G. Polson. Diagnostic measures for model criticism. *Journal of the American Statistical Association*, 91(434):753–762, 1996. 29

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. 19

A. Criminisi and J. Shotton. *Decision Forests for Computer Vision and Medical Image Analysis*. Springer Publishing Company, Incorporated, 2013. 122

B. Dai, B. Xie, N. He, Y. Liang, A. Raj, M. Balcan, and L. Song. Scalable kernel methods via doubly stochastic gradients. In *NIPS*, pages 3041–3049, 2014. 126

J. Dauxois and G. M. Nkiet. Nonlinear canonical analysis and independence tests. *The Annals of Statistics*, 26(4):1254–1278, 1998. 64

R. M. Dudley. *Real analysis and probability*, volume 74. Cambridge University Press, 2002. 21, 26

G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *UAI*, pages 258–267, 2015. 30

T. Epps and K. Singleton. An omnibus test for the two-sample problem using the empirical characteristic function. *Journal of Statistical Computation and Simulation*, 26(3–4):177–203, 1986. 33, 94

S. M. A. Eslami, D. Tarlow, P. Kohli, and J. Winn. Just-In-Time Learning for Fast and Flexible Inference. In *NIPS*, pages 154–162, 2014. 15, 116, 120, 121, 122, 123, 128, 129, 130, 132, 139, 140

A. Feuerverger. A consistent test for bivariate dependence. *International Statistical Review*, 61 (3):419–433, 1993. 64

J. Frank J. Massey. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951. 93

M. Fromont, B. Laurent, M. Lerasle, and P. Reynaud-Bouret. Kernels based tests with non-asymptotic bootstrap approaches for two-sample problems. In *COLT*, pages 23.1–23.22, 2012. 30

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *NIPS*, pages 489–496, 2008. 22, 64

A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1:1–19, 2006. 115

P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1): 3–42, Apr. 2006. 140

L. J. Gleser. On a measure of test efficiency proposed by R. R. Bahadur. 35(4):1537–1544, 1964. 101, 102

L. J. Gleser. The comparison of multivariate tests of hypothesis by means of Bahadur efficiency. 28(2):157–174, 1966. 101, 102, 103

N. Goodman, V. Mansinghka, D. Roy, K. Bonawitz, and J. Tenenbaum. Church: A language for generative models. In *UAI*, pages 220–229, 2008. 115

J. Gorham and L. Mackey. Measuring sample quality with Stein's method. In *NIPS*, pages 226–234, 2015. 94, 96

J. Gorham and L. Mackey. Measuring sample quality with kernels. In *ICML*, pages 1292–1301. PMLR, 2017. 94, 98

A. Gretton. A simpler condition for consistency of a kernel independence test. Technical report, 2015. URL `http://arxiv.org/abs/1501.06103`. 66, 67

A. Gretton and L. Györfi. Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, 11:1391–1423, 2010. 65

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory (ALT)*, pages 63–77. 2005a. 12, 63, 65, 66, 73

A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005b. 11, 12, 24

A. Gretton, A. J. Smola, O. Bousquet, R. Herbrich, A. Belitski, M. Augath, Y. Murayama, J. Pauls, B. Schölkopf, and N. K. Logothetis. Kernel constrained covariance for dependence measurement. In *AISTATS*, volume 10, pages 112–119, 2005c. 11

A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *NIPS*, pages 513–520, 2006. 12, 21

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. In *NIPS*, pages 585–592. 2008. 63

A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur. A fast, consistent kernel two-sample test. In *NIPS*, pages 673–681, 2009. 24

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012a. 11, 12, 13, 19, 20, 21, 23, 24, 26, 30, 31, 40, 93, 109

A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *NIPS*, pages 1205–1213, 2012b. 13, 30, 37, 40, 41, 99, 100

L. Györfi and E. C. van der Meulen. A consistent goodness of fit test based on the total variation distance. In G. Roussas, editor, *Nonparametric Functional Estimation and Related Topics*, pages 631–645, 1990. 93

A. Habibian, T. Mensink, and C. G. Snoek. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *ACM International Conference on Multimedia*, pages 17–26, 2014. 77

N. Heess, D. Tarlow, and J. Winn. Learning to pass expectation propagation messages. In *NIPS*, pages 3219–3227. 2013. 14, 116, 120, 121, 128, 132, 139

R. Heller, Y. Heller, S. Kaufman, B. Brill, and M. Gorfine. Consistent distribution-free k-sample and independence tests for univariate random variables. *Journal of Machine Learning Research*, 17(29):1–54, 2016. 65

W. Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.*, 19(3):293–325, 09 1948. 143, 144, 145

X. Huo and G. J. Székely. Fast computing for distance covariance. *Technometrics*, 58(4):435–447, 2016. 64

F. Hutter. *Automated Configuration of Algorithms for Solving Hard Computational Problems*. PhD thesis, University of British Columbia, Department of Computer Science, Vancouver, Canada, October 2009. http://www.cs.ubc.ca/~hutter/papers/Hutter09PhD.pdf. 122

S. Janson. The asymptotic distributions of incomplete U-statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 66(4):495–505, Sep 1984. URL https://doi.org/10.1007/BF00531887. 23

W. Jitkrittum, A. Gretton, N. Heess, S. M. A. Eslami, B. Lakshminarayanan, D. Sejdinovic, and Z. Szabó. Kernel-based just-in-time learning for passing expectation propagation messages. In *UAI*, 2015. 15

W. Jitkrittum, Z. Szabó, K. P. Chwialkowski, and A. Gretton. Interpretable Distribution Features with Maximum Testing Power. In *NIPS*, pages 181–189. 2016. 65, 94, 99, 100, 109, 113

W. Jitkrittum, Z. Szabó, and A. Gretton. An Adaptive Test of Independence with Analytic Kernel Embeddings. In *ICML*. 2017. 94, 99, 100

M. R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, 2008. 61

J. Kowalski and X. M. Tu. *Modern Applied U-Statistics*. John Wiley & Sons, 2008. 143, 145

S. G. Krantz and H. R. Parks. *A Primer of Real Analytic Functions*. Springer Science & Business Media, 2002. 92

B. Lakshminarayanan, D. Roy, and Y.-W. Teh. Mondrian forests: Efficient online random forests. In *NIPS*, pages 3140–3148, 2014. 122

Q. Le, T. Sarlós, and A. Smola. Fastfood - approximating kernel expansions in loglinear time. *ICML, JMLR W&CP*, 28:244–252, 2013. 117, 123

E. L. Lehmann. *Elements of Large-Sample Theory*. Springer Science & Business Media, 1999. 145

C. Ley, G. Reinert, and Y. Swan. Stein's method for comparison of univariate distributions. *Probability Surveys*, 14:1–52, 2017. 94

Y. Li, K. Swersky, and R. Zemel. Generative moment matching networks. In *ICML*, pages 1718–1727, 2015. 30

M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`. 133

Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *ICML*, pages 276–284, 2016. 12, 94, 95, 96, 97, 108, 109, 110, 111

J. R. Lloyd and Z. Ghahramani. Statistical model criticism using kernel two sample tests. In *NIPS*, pages 829–837, 2015. 21, 29, 30, 93

J. R. Lloyd, D. Duvenaud, R. Grosse, J. B. Tenenbaum, and Z. Ghahramani. Automatic construction and Natural-Language description of nonparametric regression models. In *AAAI*, pages 1242–1250, 2014. 29, 30

D. Lopez-Paz, P. Hennig, and B. Schölkopf. The randomized dependence coefficient. In *NIPS*, pages 1–9. 2013. 64, 73

D. Lundqvist, A. Flykt, and A. Öhman. The Karolinska directed emotional faces-KDEF. Technical report, ISBN 91-630-7164-9, 1998. 45

C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008. 43, 48

T. Minka, J. Winn, J. Guiver, S. Webster, Y. Zaykov, B. Yangel, A. Spengler, and J. Bronskill. Infer.NET 2.6, 2014. Microsoft Research Cambridge. `http://research.microsoft.com/infernet`. 14, 15, 115, 127

T. P. Minka. *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Massachusetts Institute of Technology, 2001. `http://research.microsoft.com/en-us/um/people/minka/papers/ep/minka-thesis.pdf`. 116, 119

B. Mityagin. The zero set of a real analytic function. Dec. 2015. arXiv: 1512.07276. 27, 67, 68, 97

E. Moulines, F. R. Bach, and Z. Harchaoui. Testing for homogeneity with kernel Fisher discriminant analysis. In *NIPS*, pages 609–616. 2008. 11, 30

K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2): 1–141, 2017. 17

J. Mueller and T. Jaakkola. Principal differences analysis: Interpretable characterization of differences between distributions. In *NIPS*, pages 1693–1701, 2015. 30

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, pages 429–443, 1997. 21

C. Oates, J. Cockayne, F. Briol, and M. Girolami. Convergence rates for a class of estimators based on Stein's method, 2017a. 94

C. J. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):695–718, 2017b. 94, 96

J. Oliva, B. Póczos, and J. Schneider. Distribution to distribution regression. In *ICML*, pages 1049–1057, 2013. 116

J. Pearl. Reverend Bayes on inference engines: A distributed hierarchical approach. In *AAAI*, pages 133–136, 1982. 118

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 140

B. Poczos, A. Rinaldo, A. Singh, and L. Wasserman. Distribution-free distribution regression. *arXiv:1302.0082 [cs, math, stat]*, Feb. 2013. arXiv: 1302.0082. 24, 116

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *NIPS*, pages 1177–1184, 2007. 12, 14, 64, 73, 117, 123, 124, 125, 126, 136, 137

A. Ramdas, S. Jakkam Reddi, B. Póczos, A. Singh, and L. Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *AAAI*, pages 3571–3577, 2015. 43

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006. 117, 124

M. L. Rizzo. New goodness-of-fit tests for Pareto distributions. *ASTIN Bulletin: Journal of the International Association of Actuaries*, 39(2):691–715, 2009. 93

W. Rudin. *Fourier Analysis on Groups: Interscience Tracts in Pure and Applied Mathematics, No. 12*. Literary Licensing, LLC, 2013. 125, 136

B. Schölkopf, A. Smola, and K.-R. Müller. Kernel principal component analysis. *Artificial Neural Networks-ICANN'97*, pages 583–588, 1997. 19

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.*, 41(5):2263–2291, 2013. 30, 64, 74

R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 2009. 70, 81, 98, 143, 144, 145

J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004. 19, 22

G. R. Shorack. *Probability for statisticians*. Springer Science & Business Media, 2000. 21

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 13–31, 2007. 12, 20, 66, 138

L. Song, A. Gretton, and C. Guestrin. Nonparametric tree graphical models. *AISTATS, JMLR W&CP*, 9:765–772, 2010. 123

L. Song, A. Gretton, D. Bickson, Y. Low, and C. Guestrin. Kernel belief propagation. *AISTATS, JMLR W&CP*, 10:707–715, 2011. 123

N. Srebro and S. Ben-David. Learning bounds for support vector machines with learned kernels. In *COLT*, volume 4005, pages 169–183. Springer, 2006. 57

B. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schoelkopf, and G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11: 1517–1561, 2010. 20, 21, 22, 42, 66, 67

B. Sriperumbudur, K. Fukumizu, and G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *J. Mach. Learn. Res.*, 12:2389–2410, 2011. 12, 22, 26, 36

Stan Development Team. Stan: A C++ library for probability and sampling, version 2.4, 2014. URL `http://mc-stan.org/`. 115

I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008. 17, 18, 19, 25, 26, 56, 66, 95, 96

D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton. Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy. 2016. arXiv: 1611.04488. 13, 40, 99, 100, 109

Z. Szabó, B. Sriperumbudur, B. Póczos, and A. Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40, 2016. 25, 116

G. Székely and M. Rizzo. Testing for equal distributions in high dimension. *InterStat*, (5), 2004. 30

G. J. Székely and M. L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58–80, 2005. 93

G. J. Székely and M. L. Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, 3 (4):1236–1265, 2009. 64, 74

G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007. 64

I. Tolstikhin, B. Sriperumbudur, and K. Muandet. Minimax estimation of kernel mean embeddings. *arXiv preprint arXiv:1602.04361*, 2016. 20

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 2000. 38, 50

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000. 61, 69, 70, 114

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008. 119

H. Wang and C. Schmid. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3551–3558, 2013. 77

C. K. I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *NIPS*, pages 682–688. 2001. 13

D. Wingate, N. Goodman, A. Stuhlmueller, and J. Siskind. Nonstandard interpretations of probabilistic programs for efficient inference. In *NIPS*, pages 1152–1160, 2011. 115

Z. Yang, A. J. Smola, L. Song, and A. G. Wilson. Á la carte - learning fast kernels. In *AISTATS*, 2015. 117, 123

W. Zaremba, A. Gretton, and M. Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In *NIPS*, pages 755–763, 2013. 13, 30

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *UAI*, pages 804–813, 2011. 64

Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, pages 1–18, 2017. 12, 13, 64, 73, 74, 94

J. Zhao and D. Meng. FastMMD: Ensemble of circular discrepancy for efficient two-sample test. *ArXiv e-prints*, May 2014. 12